COMPENDIUM SERIES
in Mathematics, Physics and Informatics
Vol. 4, 2014

Výber z prác
Pavol Brunovský

EDITOR
Pavel P. Povinec

UNIVERZITA KOMENSKÉHO
Fakulta matematiky, fyziky a informatiky
Bratislava

Editoriál

Fakulta matematiky, fyziky a informatiky na Univerzite Komenského v Bratislave začala v roku 2010 projekt vydávania Compendium Series in Mathematics, Physics and Informatics. Séria zborníkov prác je zameraná na publikovanie súborov vybraných prác významných osobností fakulty s cieľom zhrnúť ich úspechy a sprístupniť ich mladej generácii vedcov a študentov.

Štvrtý zväzok je venovaný dielu prof. RNDr. Pavla Brunovského, DrSc., známeho slovenského matematika a zakladateľa slovenskej matematickej školy optimálneho riadenia a dynamických systémov.

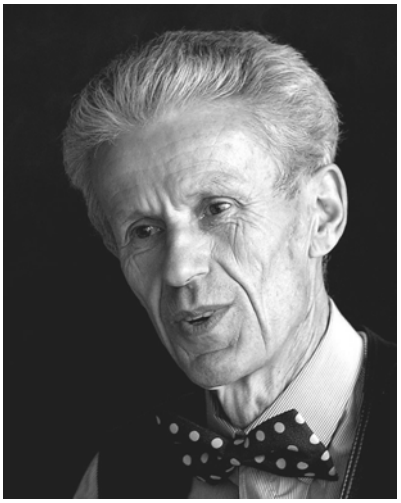
Profesor Brunovský dosiahol celý rad významných výsledkov, akými sú napríklad regularita syntézy optimálneho riadenia pre dôležité triedy úloh, alebo kánonická forma pre lineárne riadené systémy, často tiež nazývaná Brunovského normálna forma. Významné výledky dosiahol tiež v oblasti klasifikácie typických bifurkácií jednoparametrických diskrétnych dynamických systémov. Založil a viedol seminár z kvalitatívnej teórie dynamických systémov, v ktorom vyrástlo viacero matematických osobností známych aj v zahraničí. Vďaka jeho iniciatíve úsiliu vznikol a dnes sa naďalej rozvíja na Fakulte matematiky, fyziky a informatiky UK študijný program Ekonomická a finančná matematika.

Dúfame, že tento zborník vedeckých a publicistických prác bude užitočným príspevkom k udržaniu inštitucionálnej pamäti a napomôže mladým kolegom a študentom v ich ďalšom štúdiu a vedeckej práci.

Pavel P. Povinec
Editor

Obsah

Životopis a dielo profesora Pavla Brunovského



Profesor Pavel Brunovský

Profesor Pavel Brunovský sa narodil 5. decembra 1934 vo Viedni. Štúdium matematiky ukončil v roku 1958 na Prírodovedeckej Fakulte Univerzity Komenského v Bratislave. V rokoch 1969-1970 pracoval na Ústave technickej kybernetiky Slovenskej akadémie vied, kde absolvoval doktorandské štúdium v roku 1964 pod vedením profesora J. Kurzweila z Prahy. Po roku 1970 pôsobil na Matematickom ústave Slovenskej akadémie vied až do roku 1974. Najvyššiu vedeckú hodnosť, DrSc., získal v roku 1978 a za profesora bol vymenovaný v roku 1991. Dodnes aktívne pôsobí na Fakulte matematiky, fyziky a informatiky Univerzity Komenského. Bol hosťujúcim profesorom na univerzite vo Florencii, viackrát na Michiganskej štátnej univerzite i na univerzitách vo Viedni, v Tokiu, Nice a Paríži. V roku 2005 mu bol prezidentom Slovenskej republiky udelený Pribinov kríž za zásluhy v oblasti rozvoja vedy a vzdelávania.

Na začiatku vedeckej kariéry Pavla Brunovského bola hlavnou oblasťou

jeho záujmu teória a aplikácie optimálneho riadenia. Dosiahol celý rad pozoruhodných a významných výsledkov, akými sú napríklad regularita syntézy optimálneho riadenia pre dôležité triedy úloh, alebo kánonická forma pre lineárne riadené systémy, dnes nazývaná Brunovského normálna forma. Dôkazom uznania bola aj pozvaná prednáška na Svetový kongres matematikov v roku 1978, kde predniesol svoje výsledky o štruktúre optimálnej spätnej väzby. O teórii optimálneho riadenia a riadených sústav napísal tri knihy, jednu z nich v spolupráci s Jánom Černým a ďalšiu potom neskôr spoločne s Margarétou Halickou a Pavlom Jurčom.

Neskôr upriamil svoju pozornosť na príbuznú oblasť matematického výskumu a onedlho sa stal expertom aj v teórii dynamických systémov. Významné výsledky dosiahol v oblasti klasifikácie typických bifurkácií jednoparametrických diskrétnych dynamických systémov. Na tieto výsledky potom nadviazal ďalšími prácami z oblasti teórie bifurkácií konečnorozmerných dynamických systémov a chaotickej dynamiky parciálnych diferenciálnych rovníc prvého rádu.

V 80-tych rokoch sa začal venovať kvalitatívnej teórii evolučných parciálnych diferenciálnych rovníc. Využil pri tom svoje bohaté skúsenosti z teórie dynamických systémov a ich bifurkácii. Skúmal a charakterizoval kompaktné globálne atraktory priestorovo homogénnych reakčno-difúznych rovníc v jednorozmernom priestore. V ďalších prácach sa venoval skúmaniu generickosti Morseovej-Smaleovej vlastnosti pre istú triedu evolučných parciálnych diferenciálnych rovníc parabolického typu. Založil a viedol seminár z kvalitatívnej teórie dynamických systémov, v ktorom vyrástlo viacero matematických osobností medzinárodného formátu.

Na seminári z diferenciálnych rovníc v roku 1987.

V roku 1994 sa začal venovať príprave nového študijného programu Ekonomická a finančná matematika, ktorý predstavoval jedinečný program a príležitosť pre študentov aplikovať získané matematické poznatky v odboroch matematickej ekonómie a teórie financií. Vďaka jeho iniciatíve a úsiliu vznikol a dnes sa naďalej rozvíja úspešný štúdijný program Ekonomická a finančná matematika. Program založil v období, keď počet matematicky orientovaných študentov klesal nielen na Univerzite Komenského, ale všeobecne na území celého Slovenska. Program je populárnou možnosťou pre praxou motivované štúdium matematiky a zároveň je vysoko cenený talentovanými študentmi kvôli kvalite teoretického vzdelávania, ktorú ponúka. Pavel si plne uvedomoval potrebu vedeckého výskumu, ktorý by podporoval neustále zlepšovanie kvality pedagogického procesu na novozaloženom študijnom programe. Preto obohatil svoje výskumné záujmy o problémy súvisiace s modelovaním ekonomických a finančných procesov. Svoje vedomosti z kvalitatívnej teórie diferenciálnych rovníc zužitkoval vo vedeckých článkoch venovaných analýze fluktuácií na finančných trhoch ako aj v modeloch optimálneho zaistenia finančných portfólií.

Stretnutie s absolventmi odboru EFM v roku 2008

Významnou súčasťou vedecko-pedagogického pôsobenia profesora Pavla Brunovského bola aj popularizačná a publicistická aktivita. V sérii článkov propagoval svoje zaujímavé pohľady na rôzne matematické problémy. V popularizačných matematických článkoch sa najmä mladým čitateľom snažil priblížiť pohľad skúseného matematika na riešenie zaujímavých úloh, väčšinou pochádzajúcich z

rôznych aplikácií. Jeho publicistická aktivita je enormná. Svoje úvahy o postavení a úlohách matematiky a matematikov v modernej spoločnosti a celkovom smerovaní vedy na Slovensku publikoval v mienkotvorných časopisoch. Jeho kritické názory na chod spoločnosti a vedy si čitatelia mohli prečítať nielen v súčasnej postmodernej dobe, ale aj v časoch reálneho socializmu, keď písať o citlivých a často provokujúcich témach vôbec nebolo jednoduché.

Životný elán, fyzická a duševná kondícia a elán do práce sú integrálnou súčasťou osobnosti Pavla Brunovského. Je trojnásobným majstrom Československa v orientačnom behu, majstrom západoslovenského kraja a akademickým majstrom Slovenska v štafete v behu na lyžiach. Dodnes rekreačne športuje a je pre mladších kolegov príkladom aktívneho prístupu k životu v duchu gréckej *Kalokagathie*, teda harmonického súladu a vyváženosti duševnej a telesnej sily.

Vedecké dielo Pavla Brunovského svedčí o jeho obrovskom nasadení a vplyve na vedecký život na Slovensku. Z jeho osobnosti a práce vyžaruje množstvo pozitívnej energie, optimizmu a nadšenia. Pre mladších kolegov, ktorí sa mali možnosť s profesorom Pavlom Brunovským stretávať, diskutovať a či riešiť rôzne teoretické i aplikované problémy, to vždy bola a stále je inšpirujúca a povzbudzujúca skúsenosť. *Ad multos annos, sanos, fortunatos et beatos!*

Daniel Ševčovič

Kompletný zoznam vedeckých, odborných a publicistických prác profesora Pavla Brunovského

Vedecké práce

[1] P. Brunovský: *O zovšeobecnených algebrických systémoch.* Acta Facultatis Rerum Naturalium Universitatis Comeniana: Mathematica, Vol. 3, 1958 S. 41-54.

[2] P. Brunovský: *O určovaní tlmenia z grafu.* Strojnícky časopis, Roč. 11 (1961), s. 194-200.

[3] P. Brunovský: *O Emdenovej-Fowlerovej rovnici v prípade* $n < 1$. Matematicko-fyzikálny časopis SAV, Roč. 12, č. 1 (1962), s. 60-80.

[4] P. Brunovský: *A condition of the existence of a universal best* ε*-stabilizing control.* Czechoslovak Mathematical Journal, Vol. 15, No. 3 (1965), s. 370-377.

[5] P. Brunovský: *Ob analitičeskom konstruktirovanii reguľatorov s nekvadratičeskim minimizirujemym funkcionaľom.* Časopis pro pěstování matematiky, Roč. 90, č. 3 (1965), s. 290-310.

[6] P. Brunovský: *On the best stabilizing control under a given class of perbutations.* Czechoslovak Mathematical Journal, Vol. 15, No. 3 (1965), s. 329-369.

[7] P. Brunovský: *A bang-bang principle in the problem of ε-stabilization of linear control systems.* Časopis pro pěstování matematiky, Roč. 91, č. 3 (1966), s. 344-351.

[8] P. Brunovský: *O stabilizacii linejnych sistem pri opredelennom klasse postojanno dejstvujuščich vozmučšenij.* Differenciaľnye Uravnenia, Vol. 2 (1966), s. 769-777.

[9] P. Brunovský: *On the stabilization of linear systems under persitent perturbations.* 3rd Congress of International Federation of Automatic Control (IFAC): Proceedings, London: Butterwoths, 1966 40G, S. 10.

[10] P. Brunovský: *Über das schnellste Suchen eines Punktes auf einer Linie.* Matematicko-fyzikálny časopis SAV, Roč. 16 (1966), s. 97-104.

[11] P. Brunovský: *On optimal stabilization of nonlinear systems.* Mathematical Theory of Control, London: Academic Press, 1967 S. 180-189.

[12] P. Brunovský: *On optimal stabilization of periodic motions.* 4th Conference on Nonlinear Oscilation, Praha: Academia, 1968 S. 127-130.

[13] P. Brunovský: *On the necessity of a certain convexity condition for lower closure of control problems.* SIAM Journal on Control and Optimization, Vol. 6, No. 2 (1968), s. 174-185.

[14] P. Brunovský: *On the optimal stabilization of nonlinear system.* Czechoslovak Mathematical Journal, Vol. 18 (1968), s. 278-293.

[15] P. Brunovský: *Controllability and closed-loop linear controls in linear periodic system.* Journal of Differential Equations, Vol. 6, No. 2 (1969), s. 296-313.

[16] P. Brunovský: *A classification of linear controllable systems.* Kybernetika, Vol. 6, No. 3 (1970), s. 173-188.

[17] P. Brunovský: *On one-parameter families of diffeomorphisms.* Commentationes Mathematicae Universitatis Carolinae, Vol. 11, No. 3 (1970), s. 559-582.

[18] P. Brunovský: *Scorza-Dragoni's theorem for set-valued functions with unbouded values and its applications to control problems.* Matematický časopis SAV, Roč. 20, č. 3 (1970), s. 205-214.

[19] P. Brunovský: *The existence of the best stabilizing control in higher dimensions.* Mathematical Systems Theory, Vol. 4, No. 1 (1970), s. 1-5.

[20] P. Brunovský: *A concept of invariance and attractivity for multivalued differential equations.* Differential Games and Related Topics, Amsterdam: North-Holland, 1971 S. 201-208.

[21] P. Brunovský: *On one-parameter families of diffeomorphisms II: Generic branching in higher dimensions.* Commentationes Mathematicae Universitatis Carolinae, Vol. 12, No. 4 (1971), s. 765-784.

[22] P. Brunovský: *One-parameter families of diffeomorphisms.* Symposium on Differential Equations and Dynamical Systems, New York: Springer, 1971 S. 29-33.

[23] P. Brunovský: *Reprezentácia zobrazenia vstup-výstup pomocou dynamického systému s lineárnou dynamikou.* Kybernetické aspekty problémov identifikácie, Bratislava: UTK SAV, 1971 S. 85-90.

[24] P. Brunovský, Y. Alekal, Dong H. Chyung, E. B. Lee: *The quadraic problem for systems with time delays.* IEEE Transactions on Automatic Control, Vol. 16, No. 6 (1971), s. 673-687.

[25] P. Brunovský: *Generic one-parameter flows on the torus and bifurcation of periodic orbits.* EQUADIFF 3: Proceedings, Brno: J. E. Purkyně University, 1973 S. 99-104.

[26] P. Brunovský, J. Černý: *A note on information without probability.* Information and Control (Shenyang), Vol. 25 (1974), s. 134-144.

[27] P. Brunovský: *Generic properties of the rotation number of one - parameter diffeomorphisms of the circle.* Czechoslovak Mathematical Journal, Vol. 24, No. 1 (1974), s. 74-90.

[28] P. Brunovský: *On the completion of linear differential games by state-dependent strategies.* Kybernetika, Vol. 10, No. 1 (1974), s. 1-12.

[29] P. Brunovský: *The closed-loop time-optimal control I: Optimality.* SIAM Journal on Control and Optimization, Vol. 12, No. 4 (1974), s. 624-634.

[30] P. Brunovský, S. Mirica0: *Classical and Fillipov solutions of the differential equations defined by feedback control.* Revue Roumaine de Mathématiques Pures et Appliquées, Vol. 20 (1975), s. 873-883.

[31] P. Brunovský, C. Lobry: *Controlabilité bang-bang, controlabilité differentiable et perturbations des system non-linéaires.* Annali di Matematica Pura ed Applicata, Vol. 105, No. 1 (1975), s. 93-119.

[32] P. Brunovský, A. Brunovská, J. Ilavský: *Identifikácia difúzneho koeficienta.* Sympózium SKS o aplikáciách teoretických princípov kybernetiky, Bratislava: ÚTK SAV, 1976 S. 478-484.

[33] P. Brunovský: *Local controllability of odd systems.* Mathematical Control Theory, Warszawa: PWN, 1976 S. 39-46.

[34] P. Brunovský, J. Černý: *Matematická teória systémov-mýty a skutočnosť.* Sympózium SKS o aplikáciách teoretických princípov kybernetiky: Zborník, Bratislava: ÚTK SAV, 1976 S. 1-32.

[35] P. Brunovský: *Stavová teória nelineárnych systémov.* Zborník konferencie SVTS o ASRTP, Banská Bystrica: SVTS, 1976 S. 22-25.

[36] P. Brunovský: *The closed-loop time-optimal control II: Stability.* SIAM Journal on Control and Optimization, Vol. 14, No. 1 (1976), s. 156-162.

[37] P. Brunovský, A. Brunovská, J. Ilavský: *Estimation of the diffusion coefficient from sorbtion measurements.* Chemické listy, Roč. 32 (1977), s. 717-722.

[38] P. Brunovský: *Every normal linear system has a regular time-optimal synthesis.* Mathematica Slovaca, Vol. 28, No. 1 (1978), s. 81-100.

[39] P. Brunovský, A. Brunovská: *Optimal temperature control of a stirred adsorber.* Chem. Eng. Sci, Vol. 34 (1979), s. 379-386.

[40] P. Brunovský, J. Ilavský, J, Valtýni, J. Vanko: *Optimalizácia ohrevu polymerizačného reaktora.* Chem. Prumysl 29/54 (1979), 119-124.

[41] P. Brunovský: *Existence of regular synthesis for general control problems.* Journal of Differential Equations, Vol. 38, No. 3 (1980), s. 317-343.

[42] P. Brunovský: *On the structure of optimal feedback systems.* Proceedings of the International Congress of Mathematicians, Helsinki 1978, Helsinki: Academia Scientarium Fennica, 1980 S. 842-846.

[43] P. Brunovský: *Regular synthesis and singular extremals.* Optimization Techniques: Proceedings, Vol. 1, Berlin: Springer, 1980 S. 280-284.

[44] P. Brunovský: *Regular synthesis for the linear quadratic optimal control problem with linear control constraints.* Journal of Differential Equations, Vol. 38, No. 3 (1980), s. 344-360.

[45] P. Brunovský, I. Ilavský, M. Králik: *Constructing models of flow chemical technology systems by realization theory.* Chemical papers - Chemické zvesti, Vol. 35, No. 3 (1981), s. 298-312.

[46] P. Brunovský, J. Komorník: *The Riccati equation solution on the linear-Quadratic problem with constrained terminal state.* IEEE Transactions on Automatic Control, Vol. 26, No. 2 (1981), s. 398-402.

[47] P. Brunovský, J. Komorník, V. Varhola: *Adaptívny algoritmus riadenia bieliacej linky.* ASRTP 82: Zborník prednášok z 5. celoštátnej konferencie, Žilina: ČSVTS, 1982 S. 126-176.

[48] P. Brunovský, P. Mederly, H. Mederlyová, P. Meravý: *Dynamický model vodovodnej siete pre ASR vodárenskej prevádzky.* ASRTP 82: Zborník prednášok z 5. celoštátnej konferencie, Žilina: ČSVTS, 1982 S. 335-342.

[49] P. Brunovský, A. Brunovská: *Optimal temperatures for the periodic steady state for a cascade of stirred adsorbers.* Collection of Czechoslovak Chemical Communications, Vol. 47 (1982), s. 899-900.

[50] P. Brunovský, M. Otto, M. Šnejdárková, A. Ottová-Leitmannová, J. Gažo: *Model of arginine dynamics in the Japanese quail 2: Development and application of a mathematical model of argnine dynamics.* Nutrition Reports International, Vol. 28 (1983), s. 761-771.

[51] P. Brunovský: *Notes on chaos in the cell population partial differential equation.* Nonlinear Analysis-Theory, Methods & Applications, Vol. 7, No. 2 (1983), s. 167-176.

[52] P. Brunovský, J. Komorník: *The matrix Riccati equation and the noncontrollabe linear-quadratic problem with terminal constraints.* SIAM Journal on Control and Optimization, Vol. 21, No. 2 (1983), s. 280-288.

[53] P. Brunovský, P. Mederly, H. Mederlyová, P. Meravý: *DYNMOD-Program pre dynamické modelovanie vodovodnej siete.* Vodní hospodářství, Roč. 34, č. 8, (1984), s. 201-204.

[54] P. Brunovský, J. Komorník : *Ergodicity and exactness of the shift on* $C[0,1]$ *and the dynamics of a first order partial differential equation.*

Journal of Mathematical Analysis and Applications, Vol. 104, No. 1 (1984), s. 235-245.

[55] P. Brunovský, Shui-Nee Chow: *Generic properties of stationary state solutions of reaction-diffusion equations.* Journal of Differential Equations, Vol. 53, No. 1 (1984), s. 1-23.

[56] P. Brunovský, J. L. Willems, V. Kučera: *On the assigment of invariant factor by time varying feedback strategies.* System & Control Letters, Vol. 5, No. 2 (1984), s. 75-80.

[57] P. Brunovský, P. Meravý: *Solving systems of polynomial equations by bounded and real homotopy.* Numerische Mathematik, Vol. 43, No. 3 (1984), s. 397-418.

[58] P. Brunovský, P. Meravý: *Optimizing the cost of water pipeline lying.* Mathematical Methods in Operation Research, Sofia: University, 1985 S. 1-12.

[59] P. Brunovský, J. Mallet-Paret: *Switchings of optimal controls and the equation* $y^{(4)} + |y|^{\alpha}\, sign\, y = 0$*,* $0 < \alpha < 1$. Časopis pro pěstování matematiky, Roč. 110, č. 3 (1985), s. 302-313.

[60] P. Brunovský, J. Komorník: *Explicit definition of an exact measure for the semiflow of a first order partial differential equation.* Časopis pro pěstování matematiky, Roč. 111, č. 1 (1986), s. 48-53.

[61] P. Brunovský, B. Fiedler: *Connections in scalar reaction-diffusion equations with Neumann boundary conditions.* In: Springer Lecture Notes in Math. Vol 1192 (1986), 123-128.

[62] A. Brunovská, P. Brunovský, J. Markoš, B. Remiarová: *Optimalizácia tepelného režimu adiabatického reaktora.* Ropa a uhlie 28(1986), 275-282.

[63] P. Brunovský, B. Fiedler: *Numbers of zeros on invariant manifolds in reaction- diffusion equations.* Nonlinear Analysis-Theory, Methods & Applications, Vol. 10, No. 2 (1986), s. 179-193.

[64] P. Brunovský, B. Fiedler: *Simplicity of zeros in scalar parabolic equations.* Journal of Differential Equations, Vol. 62, No. 2 (1986), s. 237-241.

[65] P. Brunovský, J. Komorník: *Dynamics of a first order PDE modelling a self-reproducing cell population.* Dynamical Systems and Enviromental Models, Berlin: Academia Verlag, 1987 S. 207-214.

[66] P. Brunovský, P. Poláčik: *Generic hyperbolicity for reaction - diffusion equations on symmetric domains.* Zeitschrift für Angewandte Mathematik und Physik, Vol. 38, No. 2 (1987), s. 172-183.

[67] P. Brunovský, B. Fiedler: *Heteroclinic connections of stationary solutions of scalar reaction-diffusion equations.* Partial Differential Equations, Warsaw: Stefan Banach International Mathematical Center, 1987 S. 39-47.

[68] P. Brunovský, T. Kmeť: *The nitrogen transformation cycle in water.* Dynamical Systems and Enviromental Models, Berlin: Academia Verlag, 1987 S. 132-138.

[69] P. Brunovský, B. Fiedler: *Connecting orbits in scalar reaction-diffusion equations.* Dynamics Reported: Expositions in Dynamical Systems, Vol. 1 (1988), s. 57-89.

[70] P. Brunovský, B. Fiedler: *Connecting orbits in scalar reaction diffusion equations II: the complete solution.* Journal of Differential Equations, Vol. 81, No. 1 (1989), s. 106-135.

[71] P. Brunovský, A. Brunovská, M. Morbidelli: *Optimal catalyst pellet activity distributions for deactivating systems.* Chem. Eng. Sci., Vol. 45, Nr. 4 (1989), s. 917-925.

[72] P. Brunovský: *The attractor of the scalar reaction-diffusion equation in a smooth graph.* Journal of Dynamics and Differential Equations, Vol. 2, No. 3 (1989), s. 293-323.

[73] P. Brunovský: *The maximal attractor of the scalar reaction-diffusion equation.* Differential Equations, New York: Dekker, 1989 S. 93-98.

[74] P. Poláčik, P. Brunovský, X. Mora, J. Sola-Morales: *Asymptotic behavior of semilinear elliptic equations on an unbounded strip.* Acta Mathematica Universitatis Comenianae-New Series, 60 (1991) 2 163-184.

[75] P. Brunovský: *Controlling the dynamics of scalar reaction diffusion equations by finite dimensional controllers.* Modelling and Inverse Problems of Control for Distributed Parameters Systems, Berlin: Springer 154 (1991) 372-375.

[76] P. Brunovský, I. Tereščák: *Regularity of invariant manifolds.* Journal of Dynamics and Differential Equations, 3 (1991) 3 313-338.

[77] P. Brunovský, P. Poláčik, B. Sandstede: *Convergence in general periodic parabolic equations in one space dimension.* Nonlinear Analysis-Theory, Methods & Applications, Vol. 18, No. 3 (1992), s. 209-215.

[78] P. Brunovský, M. Kubala: *A note on continuation algorithmus for periodic orbits.* Differential Equations, Dynamical Systems, and Control Science. A Festschrift in Honor of Lawrence Markus, New York: M. Dekker, 1994 S. 15-20.

[79] P. Brunovský: *Controlling nonuniqueness of local invariant manifolds.* Journal für die reine und angewandte Mathematik, Vol. 446, (1994), p. 115-135.

[80] P. Brunovský, D. Ševčovič: *Explanation of spurt for a non-Newtonian fluid by a diffusion term.* Quarterly of Applied Mathematics, Vol. 52, No. 3 (1994), s. 401-426.

[81] P. Brunovský: *Tracking invariant manifolds without differential forms.* Acta Mathematica Universitatis Comenianae-New Series, Vol. 65, No. 1, (1996), p. 23-32.

[82] P. Brunovský, P. Poláčik: *On the local structure of ω-limit sets of maps.* Zeitschrift für Angewandte Mathematik und Physik, 48 (1997) 976-986.

[83] P. Brunovský, P. Poláčik: *The Morse-Smale structure of a generic reaction-diffusion equation in higher space dimension.* Journal of Differential Equations, Vol. 135, No. 1 (1997), s. 129-181.

[84] P. Brunovský: *S-shaped bifurcation of singulary perturbed boundary value problem.* Journal of Differential Equations, Vol. 145, No. 1 (1998), s. 52-100.

[85] T. Nagylaki, J. Hofbauer, P. Brunovský: *Convergence of multilocus systems under weak epistasis of weak selection.* J Math Biology, 38 (1999) 103-133.

[86] P. Brunovský: *C^r-Inclination theorems for singulary perturbed equations.* Journal of Differential Equations, Vol. 155, No. 1 (1999), s. 133-152.

[87] D. Ševčovič, M. Halická, P. Brunovský: *DEA analysis for a large structured bank branch network.* Central European Journal of Operations Research, Vol. 9, No. 4 (2001), s. 329-342.

[88] P. Brunovský, G. Raugel: *Genericity of the Morse-Smale property for damped wave equations.* Journal of Dynamics and Diffierential Equations, Vol. 15, No. 2/3 (2003), p. 571-658.

[89] P. Brunovský, A. Edélyi, H.-O. Walther: *On a model of a currency exchange rate-local stability and periodic solutions.* Journal of Dynamics and Diffierential Equations, Vol. 16, No. 2 (2004), s. 401-440.

[90] P. Brunovský: *Riešiteľnosť systémov rovníc CGE modelov.* Teoretické a metodologické problémy modelov vypočítateľnej všeobecnej ekonomickej rovnováhy: Zborník prác z riešenia úlohy APVT-20-039902, Bratislava: Univerzita Komenského-FMFI, 2007 S. 11-14, 55-60.

[91] P. Brunovský: *The commons game.* Ekonomický časopis, Roč. 55, č. 8 (2007), s. 811-814.

[92] P. Brunovský, M. Lapin, I. Melicherčík, J. Somorčík, D. Ševčovič: *Risks due to variability of K-day extreme precipitation totals and other K-day extreme events.* Journal of Hydrology and Hydromechanics, Vol. 57, No. 4 (2009), s. 250-263.

[93] P. Brunovský, D. Ševčovič, J. Somorčík, D. Hroncová, K. Pospišilová: *Socio-economic impacts of pandemic influenza mitigation scenarios in Slovakia.* Ekonomický časopis, Roč. 57, č. 2 (2009), s. 163-178.

[94] P. Brunovský, A. Černý, M. Winkler: *A singular differential equation stemming from an optimal control problem in financial economics.* Applied Mathematics and Optimization, Vol. 68, No. 2 (2013), s. 255-274.

Knižné publikácie a učebnice

[95] P. Brunovský: *Matematická teória optimálneho riadenia.* Bratislava: Alfa, 1980.

[96] P. Brunovský, J. Černý: *Základy matematickej teórie systémov.* Bratislava: Veda, 1980.

[97] P. Brunovský: *Theory of invariant manifolds and its applications to differential equations.* Tokyo: Department of Mathematics Science University of Tokyo, 1993.

[98] M. Halická, P. Brunovský, P. Jurča: *Optimálne riadenie: Viacetapové rozhodovacie procesy v ekonómii a financiách.* Bratislava: EPOS, 2009.

Odborné a popularizačné práce

[99] P. Brunovský: *Derivácia a linearizácia.* Matematické obzory, Zv. 2, Bratislava: Alfa, 1972 S. 1-8.

[100] P. Brunovský: *Topologická klasifikácia diferenciálnych rovníc a štruktuálna stabilita.* Pokroky matematiky, fyziky a astronomie, Roč. 18, č. 5 (1973), s. 271-281.

[101] P. Brunovský: *O minimách a maximách a ich hľadaní I.* Matematické obzory, Zv. 8, Bratislava: Alfa, 1975 S. 3-50.

[102] P. Brunovský: *O minimách a maximách a o ich hľadaní II.* Matematické obzory, Zv. 9, Bratislava: Alfa, 1976 S. 229-238.

[103] P. Brunovský: *O minimách a maximách a o ich hľadaní III.* Matematické obzory, Zv. 10, Bratislava: Alfa, 1977 S. 33-41.

[104] P. Brunovský: *Vektory očami nečistého matematika.* Matematické obzory 16 (1980), 3–6.

[105] P. Brunovský, M. Medveď: *Bifurkácie negradientných dynamických systémov.* Pokroky matematiky, fyziky a astronomie, Roč. 27, č. 2 (1982), s. 74-92.

[106] P. Brunovský: δ. Matematické obzory 22 (1984), 75–80.

[107] P. Brunovský, P. Meravý: *Ako riešiť algebraické rovnice pomocou diferenciálnych rovníc.* Pokroky matematiky, fyziky a astronomie, Roč. 32, č. 5 (1987), s. 273-286.

[108] P. Brunovský: *Qualitative theory of ordinary differential equations.* In: School of Qualitative Aspects of Nonlinear Evolution Equations, International Centre of Theoretical Physics Trieste, 1991, 33-35.

[109] P. Brunovský: *Koniec chaosu?* Pokroky matematiky, fyziky a astronomie, Vol. 40, No. 5, (1995), p. 233-243.

[110] P. Brunovský a kol.: *25 rokov Fakulty matematiky, fyziky a informatiky Univerzity Komenského v Bratislava.* Bratislava: Peter Mačura - PEEM, 2005.

[111] P. Brunovský: *Potrebujeme matematiku?* Obzory matematiky, fyziky a informatiky, Roč. 32, č. 3 (2005), s. 1-10.

[112] P. Brunovský: *Tri otázky z „ulice".* Obzory matematiky, fyziky a informatiky, Roč. 34, č. 2 (2005), s. 1-7.

Publicistické práce

[113] P. Brunovský: *Jediné na svete.* Nové slovo, Roč. 16, č. 49 (1974), s. 2.

[114] P. Brunovský a kol.: *Matematika nebude kameňom úrazu.* Nové slovo, 5.5.1977.

[115] P. Brunovský: *Je načase položiť latku vyššie.* Nové slovo 29(33), 1987.

[116] P. Brunovský: *Príbeh naozaj neuveriteľný?* Nové slovo, 17.11.1988.

[117] P. Brunovský: *Kam kráčaš academia?* Národná obroda, 20.10.1990.

[118] P. Brunovský: *Parazitológia vedy.* In: Zborník konferencie „O tvořivosti ve vědě, politice a umění II". Brno, Nadace Universitas Masarykiana, 1993, 356 strán.

[119] P. Brunovský: *Orgán a veda.* Forum Scientiae, jún 1994.

[120] P. Brunovský: *„Stará" veda v ére počítačov.* Quark, február 1996.

[121] P. Brunovský: *Koľko stoja reformy?* .týždeň, Vol. 2, No. 7 (2005), p. 17.

[122] P. Brunovský, B. Uherčíková: *Rovná či zakrivená daň?* Domino fórum, Vol. 14, No. 3 (2005), p. 6.

[123] P. Brunovský: *... a sme na špici.* .týždeň 2.3.2009.

[124] P. Brunovský: *O priliehavosti matematiky.* .týždeň 21.3.2011.

[125] P. Brunovský: *Zasa raz profesori v médiách.* .týždeň 4.7.2011.

[126] P. Brunovský: *Zakliaty Kopec.* .týždeň 14.3.2013.

Výber z vedeckých, popularizačných a publicistických prác

P. Brunovský

Controllability and linear closed-loop controls in linear periodic systems

J. Differential Equations 6 (1969), 296–313.

JOURNAL OF DIFFERENTIAL EQUATIONS 6, 296–313 (1969)

Controllability and Linear Closed-loop Controls in Linear Periodic Systems*

PAVOL BRUNOVSKY

Institute of Technical Cybernetics, Slovak Academy of Sciences, Bratislava, Czechoslovakia

and

Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota 55455

Received June 19, 1968

Consider a linear control system

$$\dot{x} = A(t)\,x + B(t)u \tag{1}$$

$x = (x_1, ..., x_n) \in R^n$, $u = (u_1, ..., u_m) \in R^m$, $A(t)$ and $B(t)$ being real continuous on $(-\infty, \infty)$ $n \times n$-matrices respectively.

The system (1) is called controllable on $(-\infty, \infty)$, if to any two points $x^1, x^2 \in R^n$ and any $t_0 \in (-\infty, \infty)$ there is a $t_1 > t_0$ and a measurable control function $u(t)$, $t \in [t_0, t_1]$ such that the solution $x(t)$ of (1), $x(t_0) = x^1$ under $u = u(t)$ satisfies $x(t_1) = x^2$ (cf. [1]).

It is a remarkable property of autonomous controllable systems ($A(t) = A$, $B(t) = B$; A, B constant) that to any prescribed spectrum Σ there is a closed-loop control $u = Qx$, (Q possibly complex) such that the spectrum of the system (1) with $u = Qx$, i.e. of the system

$$\dot{x} = (A + BQ)x$$

is Σ.

This fact has been known for a long time in the case of $u \in R^1$. For $u \in R^m$, $m > 1$ it was apparently first explicitly stated by Popov (cf. [2], [3]), who proved the equivalence of the above property to complete controllability (cf. also [5], where the problem is formulated in a somewhat different way).

Recently, Wonham [4] presented another proof of it. In addition to Popov, he has proved that if Σ contains with any complex number its conjugate with the same multiplicity, Q can be chosen real.

Let us note that a similar result can be obtained easily from the transformation of [6], (cf. Corollary 2), which is of a somewhat different kind than in [4] and [5].

* This research was partly done under the support of NASA (NGR 24-005-063).

This paper is devoted to the proof of a similar property of controllable linear periodic systems, the spectrum of $A + BQ$ replaced by the characteristic multipliers of the system

$$\dot{x} = [A(t) + B(t)\,Q(t)]x. \tag{2}$$

Throughout this paper by a real-type (n-) spectrum Σ will always be meant a set of not necessarily distinct complex numbers $\sigma_1, \ldots, \sigma_n$, containing together with every complex number its complex conjugate with the same multiplicity. All other quantities occurring in this paper will be supposed to be real, unless stated otherwise.

Further, for any $r \times s$-matrix E denote $|E| = \sum_{i=1}^{r} \sum_{j=1}^{s} |e_{ij}|$, E' the transpose of E, vectors being regarded as one column matrices in this connection. By $Y(t, t_0)$ we shall denote the solution of the matrix equation

$$\dot{Y} = A(t)Y \tag{3}$$

with $Y(t_0, t_0) = I$, I being the unity matrix. $Y(t, 0)$ will be simply denoted by $Y(t)$.

If A, B are two matrices of $n \times n$ and $n \times m$ type respectively and the system $\dot{x} = Ax + Bu$ is controllable, we shall call $\langle A, B\rangle$ a controllable pair of matrices. It is well known (cf. [1]) that $\langle A, B\rangle$ is a controllable pair if and only if rank of the matrix $(B, AB, \ldots, A^{n-1}B)$ is n.

Before formulating the main theorem let us prove several auxiliary results, some of which are of interest by themselves.

PROPOSITION 1. *Let $\langle A, B\rangle$ be a controllable pair of matrices, $m \leqslant n$, and let B have rank m. Then, there are positive integers l_i, $i = 1, \ldots, m$ such that $\sum_{i=1}^{m} l_i = n$ and a nonsingular $n \times n$ matrix C such that $C^{-1}AC = D$, $C^{-1}B = G$, where*

$$D = \begin{pmatrix} D_{11}, \ldots, D_{1m} \\ \cdots\cdots\cdots \\ D_{m1}, \ldots, D_{mm} \end{pmatrix}, \qquad G = \begin{pmatrix} G_1 \\ \vdots \\ G_m \end{pmatrix},$$

D_{ij} are $l_i \times l_j$,

$$D_{ij} = \begin{pmatrix} 0 & \ldots, & 0 \\ \cdots\cdots\cdots & & \\ 0 & \ldots, & 0 \\ \alpha_{i,k_{j-1}+1} & \ldots, & \alpha_{i,k_j} \end{pmatrix}$$

if $i \neq j$,

$$D_{ii} = \begin{pmatrix} 0 & , & 1 & ,\ldots, & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & , & 0 & ,\ldots, & 1 \\ \alpha_{i,k_{i-1}+1} & , & \alpha_{i,k_{i-1}+2} & ,\ldots, & \alpha_{i,k_i} \end{pmatrix},$$

$$k_i = \sum_{\nu=1}^{i} l_\nu ,$$

G_i *are* $l_i \times m$,

$$G_i = \begin{pmatrix} 0, & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0, & \cdots & \cdots & \cdots & 0 \\ 0,\ldots, 0, & 1, & \gamma_{i,i+1} & ,\ldots, & \gamma_{im} \end{pmatrix}$$

(1 *is in the ith column*), $i, j = 1,\ldots, m$.
For the proof see [6].

COROLLARY 1. *If* $B = b$ *is* $n \times 1$, $\langle A, b\rangle$ *is a controllable pair, then there is a nonsingular matrix* C *such that* $C^{-1}AC = D$, $C^{-1}b = g$, *where*

$$D = \begin{pmatrix} 0, & 1, & 0,\ldots, & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0,\ldots, & 0, & 1 & \\ \alpha_1 & ,\ldots, & \alpha_n & \end{pmatrix}, \qquad g = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

COROLLARY 2. *Let* $\langle A, B\rangle$ *be a controllable pair of matrices. Then, to any (real type) spectrum* Σ *there is a complex (real) matrix* Q *such that* $A + BQ$ *has spectrum* Σ.

Proof. $\langle A, B\rangle$ being a controllable pair we can choose an $n \times \tilde{m}$ submatrix $\tilde{B}$ of $B(\tilde{m} \leqslant m)$, such that $\langle A, \tilde{B}\rangle$ is a controllable pair and $\tilde{B}$ has maximal rank (cf. [6], p. 771). Suppose $\tilde{B}$ consists of the first $\tilde{m}$ columns of B. By Proposition 1 we can find a matrix C such that the transformation $x = Cy$ transfers the system

$$\dot{x} = Ax + \tilde{B}u$$

to the system

$$\dot{y} = Dy + Gu$$

D and G being as in Proposition 1. Let $\lambda^n + \beta_n\lambda^{n-1} + \cdots + \beta_1$ be the polynomial with its set of roots equal to Σ. If Σ is real-type, β_i, $i = 1,\ldots, n$

are real. We define $u = Py$, where $p_{mj} = -\beta_j - \alpha_{mj}$ and recursively $p_{ij} = \delta_{k_i+1,j} - \alpha_{ij} - \sum_{v=i+1}^{m} \gamma_{iv} p_{vj}$, where δ_{ij} is the Kronecker's symbol.

Then,

$$D + GP = \begin{pmatrix} 0, 1, \ldots, 0 \\ \cdot \quad \cdot \\ 0, \ldots, 1 \\ -\beta_1, \ldots, -\beta_n \end{pmatrix}$$

and therefore its spectrum is Σ. Further, if we denote $\tilde{Q} = PC^{-1}$, then the matrix $A + \tilde{B}\tilde{Q} = C(D + GP)C^{-1}$ has also the spectrum Σ. Now, let Q be the matrix with the first $\tilde{m}$ rows equal to those of $\tilde{Q}$ and the remaining being zero. Then, $A + \tilde{B}\tilde{Q} = A + BQ$ and thus $A + BQ$ has spectrum Σ.

PROPOSITION 2. *If $\langle A, B\rangle$ is a controllable pair and B has maximal rank, then there is an $m \times n$-matrix Q such that $\langle A + BQ, b_m\rangle$ is controllable. If* $\det A > 0$, *Q can be chosen in such a way that*

$$\det\left(A + \sum_{v=i}^{m} b_v q_v'\right) > 0, \qquad i = 1, \ldots, m \tag{4}$$

where b_v are the column vectors of B and q_v are the row vectors of Q.

Proof. We can without loss of generality suppose that A, B are transformed to the special form of Proposition 1, i.e. $A = (D_{ij})$, $B = (G_i)$, $i, j = 1, \ldots, m$.

Suppose first $\det A > 0$. Denote $p_{mj} = -\alpha_{mj}$, $j = 1, \ldots, n$, $p_m = (p_{m1}, \ldots, p_{mn})'$. Then, the last row of $A + b_m p_m'$ is zero, and, consequently, $\det(A + b_m p_m') = 0$. Since $\det(A + b_m q_m') = \det A(1 + q_m' A^{-1} b_m)$ and $A^{-1} b_m \neq 0$, for any $\epsilon > 0$ we can find q_m such that $| q_m - p_m | < m^{-1} | B |^{-1} \epsilon$ and $\det(A + b_m q_m') > 0$. Further, we define recursively

$$p_{ij} = \delta_{k_i+1,j} - \alpha_{ij} - \sum_{v=i+1}^{m} \gamma_{iv} q_{vj}, \qquad p_i = (p_{i1}, \ldots, p_{in})'.$$

If

$$\det\left(A + \sum_{v=i+1}^{m} b_v q_v' + b_i p_i'\right) \neq 0,$$

we define

$$q_{ij} = \operatorname{sign} \det\left(A + \sum_{v=i+1}^{m} b_v q_v' + b_i p_i'\right) \delta_{k_i+1,j} - \alpha_{ij} - \sum_{v=i+1}^{m} \gamma_{iv} q_{vj}.$$

If

$$\det\left(A + \sum_{\nu=i+1}^{m} b_\nu q'_\nu + b_i p'_i\right) = 0,$$

from

$$\det\left(A + \sum_{\nu=i+1}^{m} b_\nu q'_\nu + b_i q'_i\right)$$

$$= \det\left(A + \sum_{\nu=i+1}^{m} b_\nu q'_\nu\right) \cdot \left[1 + q'_i \left(A + \sum_{\nu=i+1}^{m} b_\nu q'_\nu\right)^{-1} b_i\right]$$

follows that there is a q_i such that $| p_i - q_i | < m^{-1} | B |^{-1}\epsilon$ and (4) is valid (recall that q_ν, $\nu = i + 1,..., m$ are already chosen so that $A + \sum_{\nu=i+1}^{m} b_\nu q'_\nu$ is nonsingular).

From the above construction, it follows that $A + BQ = T + S$, where

$$T = \begin{pmatrix} 0, \theta_1, 0,..., 0 \\ 0, 0, \theta_2, ..., 0 \\ \cdots\cdots\cdots \\ 0,..., 0, \theta_{n-1} \\ 0,..., 0 \end{pmatrix},$$

$\theta_i = +1$ or -1 and $| S | < \epsilon$. Clearly the pair $\langle T, b_m \rangle$ is controllable. By ([7], Chapter 2, Theorem 1), for $\epsilon > 0$ sufficiently small $\langle A + BQ, b_m \rangle = \langle T + S, b_m \rangle$ is also controllable.

The assertion of the proposition for $\det A \leqslant 0$ can be obtained by a similar, rather simplified argument ($p_i = q_i$, $i = 1,..., m$).

PROPOSITION 3. *Let $A(t)$ and $B(t)$ be ω-periodic an integrable over $[0, \omega]$. Then, the system* (1) *is controllable if and only if the rows of the matrix function $Y^{-1}(s)\, B(s)$, $s \in [0, n\omega]$ are linearly independent.*

Proof. It is well known (cf. [1]) that (1) is controllable if and only if to every t_0 there is a t_1 such that the rows of the functional matrix $Y^{-1}(s)\, B(s)$ on $[t_0, t_1]$ are linearly independent. Clearly, this is true in the periodic case if and only if such a t_1 exists for every $t_0 = k\omega$, k integer. Further, we have $Y^{-1}(s, k\omega)\, B(s) = Y^{-1}(s - k\omega)\, B(s - k\omega) = Y^{-1}(t)\, B(t)$ for $s \in [k\omega, t_1]$ and $t \in [0, t_1 - k\omega]$. Consequently (1) is controllable if and only if the rows of $Y^{-1}(t)\, B(t)$ are linearly independent on $[0, t_1]$ for some $t_1 > 0$. It remains to prove that this is equivalent with linear independence of the rows of $Y^{-1}(t)\, B(t)$ on $[0, n\omega]$; the only nontrivial part of this statement is that the

linear independence of the rows of the matrix $Y^{-1}(t)\,B(t)$ on $[0, t_1]$ for $t_1 > n\omega$ implies their independence on $[0, n\omega]$. To prove this, note that from $c'Y^{-1}(t)\,B(t) \equiv 0$ on $[0, n\omega]$, $Y^{-1}(t)\,B(t) = Y^{-k}(\omega)\,Y^{-1}(t - k\omega)\,B(t - k\omega)$ and the fact that for $k \geqslant n$ the matrix $Y^{-k}(\omega)$ is a linear combination of the matrices $Y^{-i}(\omega)$, $i = 0, 1,..., n - 1$ follows $c'Y^{-1}(t)\,B(t) \equiv 0$ on any interval $[k\omega, (k + 1)\omega]$, $k \geqslant n\omega$.

COROLLARY 4. *Let $A(t), B(t)$ be continuous and ω-periodic. Then,* (1) *is controllable if and only if there are $r (1 \leqslant r \leqslant n)$ numbers*

$$0 \leqslant t_1 < \cdots < t_r < \omega \tag{5}$$

and integers $i_1,..., i_r$, $1 \leqslant i_j \leqslant m$, such that

(i) *The vectors $\tilde{b}_1 = Y^{-1}(t_1)\,b_{i_1}(t_1),..., \tilde{b}_r = Y^{-1}(t_r)\,b_{i_r}(t_r)$ are linearly independent (b_{i_j} denotes the i_jth column of B)*

(ii) *The pair of matrices $\langle Y(\omega), \tilde{B}\rangle$ is controllable, where $\tilde{B} = (\tilde{b}_1,..., \tilde{b}_r)$.*

Proof. Suppose (1) is controllable. From the set $\{Y^{-1}(t)\,b_i(t) \mid t \in [0, \omega],\ i = 1,..., m\}$ choose an arbitrary maximal set of linearly independent points $\tilde{b}_j = Y^{-1}(t_j)\,b_{i_j}(t_j)$, $0 \leqslant t_1 \leqslant \cdots \leqslant t_r \leqslant \omega$. Then (i) is satisfied and every point $Y^{-1}(t)\,b_i(t)$, $t \in [0, \omega]$ $i = 1,..., m$, is a linear combination of $\tilde{b}_j$, $j = 1,..., r$. Now, let $t \in [0, n\omega]$, $t = \tau + \mu\omega$, $\tau \in [0, \omega]$, $1 \leqslant i \leqslant m$. Then, there is an m-vector d such that $Y^{-1}(\tau)\,b_i(\tau) = \tilde{B}d$. We have

$$Y^{-1}(t)\,b_i(t) = Y^{-\mu}(\omega)\,Y^{-1}(\tau)\,b_i(\tau) = Y^{-\mu}(\omega)\,\tilde{B}d.$$

Consequently, the linear hull of the set of vectors $\{Y^{-1}(t)\,b_i(t) \mid t \in [0, \omega],\ i = 1,..., m\}$ is contained in the linear hull of the vectors $\{Y^{-\mu}(\omega)\,\tilde{b}_i \mid i = 1,..., r, \mu = 0,..., n - 1\}$. But the linear independence of the rows of the matrix $Y^{-1}(t)\,B(t)$ on $[0, n\omega]$ implies that the vectors $\{Y^{-1}(t)\,b_i(t) \mid t \in [0, n\omega],\ i = 1,..., m\}$ span R^n. Thus, the vectors $\{Y^{-\mu}(\omega)\tilde{b}_i \mid i = 1,..., r, \mu = 0,..., n - 1\}$ span R^n, or, equivalently, rank $(\tilde{B}, Y^{-1}(\omega)\,\tilde{B},..., Y^{-n+1}(\omega)\tilde{B}) = n$. Since $Y(\omega)$ is nonsingular, this is equivalent with rank $(\tilde{B}, Y(\omega)\,\tilde{B},..., Y^{n-1}(\omega)\tilde{B}) = n$.

The numbers $t_1,..., t_r$ are not necessarily distinct, but since $Y^{-1}(t)\,B(t)$ are continuous, a sufficiently small change of the numbers t_i will not affect the rank of the matrices $\tilde{B}$ and $(\tilde{B}, Y(\omega)\,\tilde{B},..., Y^{n-1}(\omega)\tilde{B})$. Therefore, by a small change of the numbers t_i we can achieve that both (5) and (i), (ii) will be valid.

In the other direction, the corollary is obvious.

Remark 1. Since $Y(\omega)$ is nonsingular, $\langle Y(\omega), \tilde{B}\rangle$ is a controllable pair if and only if $\langle Y(\omega), Y(\omega)\tilde{B}\rangle$ is controllable.

PROPOSITION 4. *Let $A(t), A_k(t)$ be ω-periodic integrable matrices such that $A_k(t) \to A(t)$ for $k \to \infty$ in $L_1(0, \omega)$. Then, $Y_k(\omega) \to Y(\omega)$ for $k \to \infty$, where $Y_k(t)$ is the solution of $\dot{Y} = A_k(t)Y$ with $Y_k(0) = I$.*

Proof. We have

$$| Y_k(t)| \leqslant n + \int_0^t | A_k(s)| \, | Y_k(s)| \, ds.$$

By Gronwall's inequality

$$| Y_k(t)| \leqslant n \exp \int_0^t | A_k(s)| \, ds \leqslant n \exp \int_0^\omega | A_k(s)| \, ds \quad \text{for} \quad t \in [0, \omega].$$

Since $A_k(t)$ converge in $L_1(0, \omega)$, $\int_0^\omega | A_k(s)| \, ds$ is bounded; hence, $| Y_k(t)|$ are equibounded on $[0, \omega]$ say $| Y_k(t)| \leqslant \kappa$. Further, we have

$$\begin{aligned} | Y_k(t) - Y(t)| &\leqslant \int_0^t | A_k(s) \, Y_k(s) - A(s) \, Y(s)| \, ds \\ &\leqslant \int_0^\omega | Y_k(s)| \, | A_k(s) - A(s)| \, ds \\ &\quad + \int_0^t | A(s)| \cdot | Y_k(s) - Y(s)| \, ds \\ &\leqslant \kappa \int_0^\omega | A_k(s) - A(s)| \, ds + \int_0^t | A(s)| \, | Y_k(s) - Y(s)| \, ds. \end{aligned}$$

Applying Gronwall's inequality, we obtain $|Y_k(t) - Y(t)| \leqslant \kappa \int_0^\omega |A_k(s) - A(s)| \, ds. \exp\{\int_0^t | A(s)| \, ds\}$ which completes the proof.

THEOREM. *Let $A(t)$, $B(t)$ be ω-periodic and C^1 in t and let* (1) *be controllable. Then,*

(i) *To any real-type spectrum $\sum = \{\sigma_1, ..., \sigma_n\}$ such that $\sigma_i \neq 0$, $i = 1, ..., n$ and $\prod_{i=1}^n \sigma_i > 0$ there is an ω-periodic $m \times n$ matrix $Q(t)$ such that the characteristic multipliers of* (2) *are equal to σ_i.*

(ii) *To any real-type spectrum $\sum = \{\sigma_1, ..., \sigma_n\}$ such that $\sigma_i \neq 0$, $i = 1, ..., n$ there is a 2ω-periodic $m \times n$ matrix $Q(t)$ such that the characteristic multipliers of* (2) *(considered as 2ω-periodic system) are equal to σ_i^2.*

Moreover, both (i) *and* (ii) *are sufficient for complete controllability of* (1).

Proof. Suppose first that (1) is not controllable. Then, there is at least one nonzero n-vector c such that

$$c' Y^{-1}(t) \, B(t) = 0 \qquad \text{for all} \qquad t. \tag{6}$$

The set of all c satisfying (6) is a linear subspace of R^n invariant under the action of $Y(\omega)'$, since

$$(Y(\omega)'c)' Y^{-1}(t) \, B(t) = c' \cdot Y^{-1}(t - \omega) \, B(t - \omega) = 0 \qquad \text{for all} \qquad t.$$

Therefore, it contains at least one eigenvector of $Y(\omega)'$, i.e. there is a vector c_0 satisfying (6) such that $Y(\omega)'c_0 = \lambda c_0$. Now, let $Q(t)$ be any periodic $m \times n$ matrix and $X(t)$ be the fundamental matrix of (2) with $X(0) = I$. Using the variation of constants formula we obtain

$$c_0'X(\omega) = c_0'Y(\omega) + \int_0^\omega Y^{-1}(t)\,B(t)Q(t)\,X(t)\,dt$$
$$= \lambda c_0' + \lambda \int_0^\omega c_0'Y^{-1}(t)\,B(t)Q(t)\,X(t)\,dt = \lambda c_0'$$

Thus, λ is an eigenvalue of $X(\omega)'$ (and, thus, of $X(\omega)$) for any Q.

Now, let (1) be controllable. Choose the numbers $0 \leqslant t_1 < \cdots < t_r < \omega$ and $\{i_1, \ldots, i_n\}$ and define $\tilde{B} = (\tilde{b}_1, \ldots, \tilde{b}_r)$ as in Corollary 4.

The proof will be accomplished in several steps which we shall number for better orientation.

1°. For an arbitrary $r \times n$ matrix $Q = (q_1, \ldots, q_r)'$ and

$$0 < h \leqslant h_0 = \min_{i=1,\ldots,r} \{t_i - t_{i-1}, \omega - t_r\}$$

denote

$$Q_h(t) = \begin{cases} \dfrac{1}{h} Q^{(j)} & \text{for} \quad t \in [t_j + \nu\omega, t_j + \nu\omega + h],\ \nu \text{ integer} \\ 0 \text{ elsewhere} \end{cases}$$

where $Q^{(j)}$ is the matrix with i_jth row q_j and the remaining rows equal to zero.

Denote $X_{Q,h}(t, \tau)$ the solution of the matrix equation

$$\dot{X} = (A(t) + B(t)Q_h(t))x$$

with $X_{Q,h}(\tau, \tau) = I$. We prove that for $t \in [0, 1]$

$$X_{Q,h}(t_j + ht, t_j) = e^{tb_{i_j}(t_j)q_j'} + O(h) \tag{7}$$

locally uniformly in Q, which implies

$$X_{Q,h}(\omega, 0) = Y(\omega, t_r)\, e^{b_{i_r}(t_r)q_r'}\, Y(t_r, t_{r-1})\, e^{b_{i_{r-1}}(t_{r-1})q_{r-1}'} \cdots e^{b_{i_1}q_1'}\, Y(t_1) + O(h)$$
$$= Y(\omega) \prod_{j=1}^{r} e^{\tilde{b}_j\tilde{q}_j'} + O(h), \tag{8}$$

locally uniformly in Q, where $\tilde{q}_j = Y(t_j)'\,q_j$. This allows us to define

$$X_{Q,0}(\omega, 0) = \lim_{h\to 0} X_{Q,h}(\omega, 0) = Y(\omega) \prod_{j=1}^{r} e^{\tilde{b}_j\tilde{q}_j'}.$$

For $t \in [0, 1]$, $0 < h \leqslant h_0$ we have

$$| X_{Q,h}(t_j + ht, t_j)| \leqslant n + \int_0^t (\alpha h + \beta | q_j |)| X_{Q,h}(t_j + hs, t_j)| \, ds$$

where

$$\alpha = \max_{t \in [0, h_0]} | A(t_j + t)|, \qquad \beta = \max_{t \in [0, h_0]} | b_{i_j}(t_j + t)|$$

and, consequently, by Gronwall's inequality

$$| X_{Q,h}(t_j + ht, t_j)| \leqslant ne^{\alpha h t} \cdot e^{\beta t |q_j|} \leqslant \kappa e^{\beta |q_j|} \quad \text{for} \quad t \in [0, 1] \tag{9}$$

where κ is a constant independent of h, Q for h sufficiently small.

Further, we have for $t \in [0, 1]$, $h \in [0, h_0]$

$$\begin{aligned}
&| X_{Q,h}(t_j + ht, t_j) - e^{t b_{i_j}(t_j) q'_j} | \\
&\quad \leqslant \int_0^t |[hA(t_j + hs) + b_{i_j}(t_j + hs) \, q'_j] \, X_{Q,h}(t_j + hs, t_j) \\
&\qquad - b_{i_j}(t_j) \, q'_j e^{s b_{i_j}(t_j) q'_j} | \, ds \\
&\quad \leqslant \int_0^t | \, b_{i_j}(t_j)|| \, q_j \, || \, X_{Q,h}(t_j + hs, t_j) \\
&\qquad - e^{b_{i_j}(t_j) q'_j} | \, ds + \int_0^t h \, | \, A(t_j + hs)|| \, X_{Q,h}(t_j + hs, t_j)| \, ds \\
&\qquad + \int_0^t | \, b_{i_j}(t_j + hs) - b_{i_j}(t_j)|| \, q_j \, || \, X_{Q,h}(t_j + hs, t_j)| \, ds.
\end{aligned} \tag{10}$$

According to (9),

$$\int_0^t h \, | \, A(t_j + hs)|| \, X_{Q,h}(t_j + hs, t_j)| \, ds \leqslant | \, h \, | \, \kappa \alpha e^{\beta |q_j|} \tag{11}$$

$$\int_0^t | \, b_{i_j}(t_j + hs) - b_{i_j}(t_j \, || \, q_j \, || \, X_{Q,h}(t_j + hs, t_j)| \, ds \leqslant | \, h \, || \, q_j \, | \cdot \kappa e^{\beta |q_j|} \cdot \beta_1 \tag{12}$$

where

$$\beta_1 = \max_{t \in [t_j, t_j + h_0]} | \, \dot{b}_{i_j}(t)|$$

From (10), (11), (12) it follows

$$\begin{aligned}
&| X_{Q,h}(t_j + ht, t_j) - e^{t b_{i_j}(t_j) q'_j} | \\
&\quad \leqslant h \cdot \gamma_1(q) + \gamma_2(q) \int_0^t | \, X_{Q,h}(t_j + hs, t_j) - e^{b_{i_j}(t_j) q'_j} | \, ds
\end{aligned}$$

for $t \in [0, 1]$, where $\gamma_1(q), \gamma_2(q)$ do not depend on h for $h \in [0, h_0]$ and are locally bounded in q. Using Gronwall's inequality, we obtain

$$| X_{Q,h}(t_j + ht, t_j) - e^{tb_{i_j}(t_j)q'_j} | \leqslant h\gamma_1(q)\, e^{\gamma_2(q)} \qquad \text{for} \quad t \in [0, 1]$$

which proves (7).

2°. Let p be any vector such that $\det(I + \tilde{b}_j p') > 0$. Then, there is a real vector q such that

$$Y^{-1}(t_j)\, e^{b_{i_j}(t_j)q'}\, Y(t_j) = I + \tilde{b}_j p' \tag{13}$$

or, equivalently,

$$e^{\tilde{b}_j \tilde{q}'} = I + \tilde{b}_j p'.$$

where $\tilde{q} = Y(t_j)'q$.

We have

$$e^{\tilde{b}_j \tilde{q}'} = I + \tilde{b}_j \tilde{q}' \frac{e^{\tilde{q}'\tilde{b}_j} - 1}{\tilde{q}'\tilde{b}_j} \tag{14}$$

(Here and further we understand $(e^\xi - 1)/\xi = 1$ if $\xi = 0$).

Using (14), (13) can be rewritten as

$$\tilde{b}_j \tilde{q}' \frac{e^{\tilde{q}'\tilde{b}_j} - 1}{\tilde{q}'\tilde{b}_j} = \tilde{b}_j p' \tag{15}$$

Denote $z_1, \ldots, z_{n-1}$ arbitrary n vectors such that $z_1, \ldots, z_{n-1}, \tilde{b}_j$ form a basis in R^n. Since $1 + p'\tilde{b}_j = \det(I + \tilde{b}_j p') > 0$, we can define

$$\tilde{q}'\tilde{b}_j = \ln(1 + p'\tilde{b}_j) \tag{16}$$

$$\tilde{q}' z_\nu = \frac{p' \ln(1 + p'\tilde{b}_j)}{p'\tilde{b}_j} \cdot z_\nu, \qquad \nu = 1, \ldots, n - 1 \tag{17}$$

(again, $\xi^{-1} \ln(1 + \xi) = 1$ if $\xi = 0$). Since z_ν, $\nu = 1, \ldots, n - 1$ and $\tilde{b}_j$ form a basis, $\tilde{q}'$ is uniquely determined by (16), (17). From (16) follows

$$e^{\tilde{q}'\tilde{b}_j} - 1 = p'\tilde{b}_j \tag{18}$$

or, equivalently,

$$\tilde{q}' \frac{e^{\tilde{q}'\tilde{b}_j} - 1}{\tilde{q}'\tilde{b}_j} \tilde{b}_j = p'\tilde{b}_j\,. \tag{19}$$

From (16), (17), (18) follows

$$\tilde{q}' z_\nu = \frac{\tilde{q}' \tilde{b}_j p'}{e^{\tilde{q}' \tilde{b}_j} - 1} \cdot z_\nu$$

or, equivalently,

$$\tilde{q}' \frac{e^{\tilde{q}' \tilde{b}_j} - 1}{\tilde{q}' \tilde{b}_j} z_\nu = p' z_\nu . \tag{20}$$

Since z_ν, $\nu = 1,..., n-1$ and $\tilde{b}_j$ form a basis, from (19) and (20) follows

$$\tilde{q}' \cdot \frac{e^{\tilde{q}' \tilde{b}_j} - 1}{\tilde{q}' \tilde{b}_j} = p'. \tag{21}$$

Multiplying this equation from the left by $\tilde{b}_j$ we obtain (15).

As a consequence of this we obtain that to any set of r vectors $p_1 ,..., p_r$ such that $\det(I + \tilde{b}_j p_j') > 0$ we can find a matrix Q such that

$$X_{Q,0}(\omega, 0) = Y(\omega) \prod_{j=1}^{r} (I + \tilde{b}_j p_j'). \tag{22}$$

3°. To any $r \times n$-matrix V, $V = (v_1 ,..., v_r)'$ having the property

$$\det \left(I + \sum_{\nu=j}^{r} \tilde{b}_\nu v_\nu' \right) > 0 \quad j = 1,..., r \tag{23}$$

there are vectors $p_1 ,..., p_r$ such that

$$\prod_{j=1}^{r} (I + \tilde{b}_j p_j') = I + \sum_{j=1}^{r} \tilde{b}_j v_j' , \det(I + \tilde{b}_j p_j') > 0. \tag{24}$$

It is easy to verify that under our assumptions the vectors

$$p_j = \left(I + \sum_{\nu=j+1}^{r} \tilde{b}_\nu v_\nu' \right)^{-1'} v_j$$

solve equation (24) and we have

$$\det(I + \tilde{b}_j p_j') = \det \left[I + \tilde{b}_j v_j' \left(I + \sum_{\nu=j+1}^{r} \tilde{b}_\nu v_\nu' \right)^{-1} \right]$$

$$= \det \left(I + \sum_{\nu=j+1}^{r} \tilde{b}_\nu v_\nu' \right)^{-1} \cdot \det \left(I + \sum_{\nu=j}^{r} \tilde{b}_\nu v_\nu' \right) > 0.$$

Combining (22) and (24) we obtain that to any $r \times n$ matrix V such that (23) is satisfied there is a matrix Q such that

$$X_{Q,0}(\omega, 0) = Y(\omega) + \sum_{j=1}^{r} Y(\omega)\, \tilde{b}_j v_j' \tag{25}$$

4°. By Liouville's theorem $\det Y(\omega) = \exp \int_0^\omega \operatorname{tr} A(t)\, dt > 0$. Therefore, according to Proposition 2 and Remark 1 there is an $r \times n$ matrix V such that $\langle Y(\omega)(I + \tilde{B}V), Y(\omega)\, \tilde{b}_r \rangle$ is controllable and $\det(Y(\omega) + \sum_{\nu=j}^{r} Y(\omega)\, \tilde{b}_\nu v_\nu') > 0$, $j = 1,..., r$, which is equivalent with (23). From ([7], Chap. 2 Theorem 11) again follows that there is an $\epsilon > 0$ such that if $| Z - Y(\omega)(I + \tilde{B}V)| < \epsilon$, $| b - \tilde{b}_r | < \epsilon$ then $\langle Z, Y(\omega)b \rangle$ is also a controllable pair.

5°. From 2°, 3°, 4° follows that there is a matrix Q such that

$$X_{Q,0}(\omega, 0) = Y(\omega)(I + \tilde{B}V)$$

and that for sufficiently small $h > 0$,

$$| X_{Q,h}(\omega, 0) - X_{Q,0}(\omega, 0)| < \tfrac{1}{2}\epsilon. \tag{26}$$

Since $b_{i_r}(t)$ and $Y(\omega, t)$ are continuous, we can choose $h > 0$ so small that $| Y(\omega, t_r + h)\, b_{i_r}(t_r + h) - Y(\omega, t_r)\, \tilde{b}_r | < \epsilon$. Denote

$$R_h^\delta(t) = \begin{cases} h^{-1} Q^{(j)} \xi_\delta[h^{-1}(t - t_j)] & \text{for} \quad t \in [t_j + \nu\omega, t_j + \nu\omega + h], \nu \quad \text{integer} \\ 0 \text{ elsewhere} \end{cases}$$

where $\xi_\delta(t) = 0$ for $t = 0, 1$, $\xi_\delta(t) = 1$ for $t \in [\delta, 1 - \delta]$ $0 \leqslant \xi_\delta(t) \leqslant 1$ for $t \in [0, 1]$ and $\xi_\delta(t)$ is C^1 on $[0, 1]$. Clearly $A(t) + B(t)\, R_h^\delta(t)$ is ω-periodic, C^1 and $A(t) + B(t)\, R_h^\delta(t) \to A(t) + B(t)\, Q_h(t)$ for $\delta \to 0$ in $L_1(0, \omega)$. Thus, if we denote $W_h^\delta(t)$ the solution of the matrix equation

$$\dot{x} = (A(t) + B(t)\, R_h^\delta(t))x, \tag{27}$$

with $W_h^\delta(0) = I$, we have by Proposition 4

$$| W_h^\delta(\omega) - X_{Q,h}(\omega, 0)| < \tfrac{1}{2}\epsilon \tag{28}$$

for sufficiently small $\delta > 0$. Combining (26), and (28) we obtain

$$| W_h^\delta(\omega) - Y(\omega)(I + \tilde{B}V)| < \epsilon$$

Hence, by 4°, $\langle W_h^\delta(\omega), Y(\omega, t_r + h)\, b_{i_r}(t_r + h) \rangle$ is controllable. Since $A(t) + B(t)$, $R_h^\delta(t) = A(t)$ for $t \in (t_r + h, \omega)$, $Y(\omega, t_r + h) = W_h^\delta(\omega)$, $W_h^\delta(t_r + h)^{-1}$ and, consequently, $\langle W_h^\delta(\omega), W_h^\delta(\omega)\, W_h^\delta(t_r + h)^{-1}\, b_{i_r}(t_r + h) \rangle$ is controllable.

6°. By the above procedure we have reduced our problem to the case of $A(t)$, $B(t)$ in (1) being C^1 and the pair $\langle Y(\omega), Y(\omega, t_1)\, b_1(t_1)\rangle$ being controllable for some $t_1 \in [0, \omega)$, since the system

$$\dot{x} = \tilde{A}(t)x + B(t)u \tag{29}$$

with $\tilde{A}(t) = A(t) + B(t)\, R_h^{\delta}(t)$ and suitably re-ordered columns of B satisfies the above properties. If the matrix $\tilde{Q}(t)$ solves our problem for the system (29), then the matrix $R_h^{\delta}(t) + \tilde{Q}(t) = Q(t)$ solves the problem for the original system (1).

7°. Let us hence suppose that $A(t)$, $B(t)$ are C^1 and $\langle Y(\omega), Y(\omega)\tilde{b}\rangle$, $\tilde{b} = Y^{-1}(t_1)\, b_1(t_1)$ is controllable. Then, there is a nonsingular matrix C such that $D = C^{-1}Y(\omega)C$, $g = C^{-1}Y(\omega)\tilde{b}$ have the special form of Corollary 1. It is easy to verify that the linear change of variables $x = Cy$ transforms $Y(\omega)$ into D, $Y(\omega)\, \tilde{b}_1$ into g without changing the characteristic multipliers of the system so that we can without loss of generality assume that $Y(\omega)$, $Y(\omega)\tilde{b}$ have already this special form of D and g of Corollary 1.

Now, choose an arbitrary spectrum containing no zero element. Choose p according to Corollary 2 in such a way that the spectrum of

$$Y(\omega) + Y(\omega)\, \tilde{b}p' = Y(\omega)[I + \tilde{b}p']$$

is Σ. If $\sigma_1 \cdots \sigma_n > 0$, then $\det(I + \tilde{b}p') > 0$ and according to 2°, there is a vector q^0 such that

$$X_{q^0,0}(\omega, 0) = Y(\omega)[I + \tilde{b}p']$$

(where $X_{q,h}(t, \tau)$ stands now for $X_{Q,h}(t, \tau)$ with $Q = (q, 0, \ldots, 0)'$). If $\sigma_1 \cdots \sigma_n < 0$, then certainly $(\sigma_1 \cdots \sigma_n)^2 > 0$ and we can apply our argument for (1) considered as a 2ω-periodic system.

8°. The proof will be complete if we show that there is an $h > 0$ and a vector q such that $X_{q,h}(\omega, 0)$ is similar to $X_{q^0,0}(\omega, 0)$. This will be proved by an implicit function argument, for which we need first to prove the continuity of

$$X_{q,h}(\omega, 0), \quad \frac{\partial}{\partial q_i} X_{q,h}(\omega, 0) \quad \text{and} \quad \frac{\partial}{\partial h} X_{q,h}(\omega, 0)$$

in q and h in the right (in h) neighborhood of the point $(q^0, 0)$. Since

$$X_{q,h}(\omega, 0) = Y(\omega, t_1 + h) \cdot X_{q,h}(t_1 + h, t_1)\, Y(t_1)$$

it is obvious that if we denote $X_{q,0}(t_1 + 0, t_1) = e^{b_1(t_1)q'}$, it is sufficient to prove the continuous differentiability of $X_{q,h}(t_1 + h, t_1)$. For the sake of simplicity we shall use further the notation $X_{q,h}(t_1 + t, t_1) = Z_{q,h}(t)$, $b_1(t_1 + t) = b(t)$, $b(0) = b$, $X_{q,0}(t_1 + 0, t_1) = Z_{q,0}$.

From the definition of $Z_{q,h}(h)$ it is evident that $Z_{q,h}(h)$ is continuous in q, h for $h > 0$. From (7) it follows that it is continuous in q for $h = 0$. Therefore, the continuity of $Z_{q,h}(h)$ in q, h for $h \geqslant 0$ follows from the local uniformity of $O(h)$ in (7).

$(\partial/\partial q_i)\, Z_{q,h}(t)$ is the solution of the equation

$$\dot{Z} = [A(t_1 + t) + h^{-1}b(t)\, q']\, Z + h^{-1}b(t)\, e_i' Z_{q,h}(t) \tag{30}$$

with $Z(0) = 0$, where e_i is the vector with ith component 1 and the remaining 0. From (30) follows

$$\frac{\partial Z_{q,h}(h)}{\partial q_i} = h^{-1} \int_0^h Z_{q,h}(h)\, Z_{q,h}^{-1}(s)\, b(s)\, e_i' Z_{q,h}(s)\, ds$$
$$= \int_0^1 Z_{q,h}(h)\, Z_{q,h}^{-1}(hs)\, b(hs)\, e_i' \cdot Z_{q,h}(hs)\, ds$$

and by (7),

$$\lim_{h\to 0} \frac{\partial Z_{q,h}(h)}{\partial q_i} = \int_0^1 e^{(1-s)bq'} b e_i' e^{sbq'}\, ds = b e_i' \frac{e^{q'b} - 1}{q'b}$$
$$+ bq' b_i \frac{-e^{q'b} + 1 + q'be^{q'b}}{(q'b)^2}$$
$$= \frac{\partial}{\partial q_i}\left(I + bq' \frac{e^{q'b} - 1}{q'b}\right) = \frac{\partial}{\partial q_i} e^{bq'} = \frac{\partial}{\partial q_i} Z_{q,0}, \tag{31}$$

(where b_i stands for the ith coordinate of b) if $q'b \neq 0$; the validity of (31) can be similarly verified if $q'b = 0$.

Since the convergence in (31) is locally uniform in q, the continuity of $(\partial/\partial q_i)\, Z_{q,h}(h)$ is proved.

Now, consider for $h, k < h_0$:

$$\frac{d}{dt}[Z_{q,h}(ht) - Z_{q,k}(kt)] = [hA(t_1 + ht) + b(ht)\, q']\, Z_{q,h}(ht)$$
$$- [kA(t_1 + kt) + b(kt)\, q']\, Z_{q,k}(kt)$$
$$= [hA(t_1 + ht) + b(ht)\, q'][Z_{q,h}(ht) - Z_{q,k}(kt)]$$
$$+ [hA(t_1 + ht) - kA(t_1 + kt) + (b(ht)$$
$$- b(kt))\, q']\, Z_{q,k}(kt). \tag{32}$$

Denote $[hA(t_1 + ht) - kA(t_1 + kt) + (b(ht) - b(kt))\, q'] = \Gamma(h, k, t)$. Using the variation of constants formula, we obtain from (32)

$$Z_{q,h}(h) - Z_{q,k}(k) = \int_0^1 Z_{q,h}(h)\, Z_{q,h}^{-1}(ht)\, \Gamma(h, k, t)\, Z_{q,k}(kt)\, dt. \tag{33}$$

We have

$$\begin{aligned}\Gamma(h, k, t) &= (h-k)\,A(t_1+ht) + k[A(t_1+ht) - A(t_1+kt)] \\ &\quad + [b(ht) - b(kt)]\,q' \\ &= (h-k)\,A(t_1+ht) + k[\dot{A}(t_1+ht)(h-k)t] \\ &\quad + \dot{b}(ht)(h-k)\,t\cdot q' + \omega(h, h-k),\end{aligned} \tag{34}$$

where $\omega(h, h-k) = o(h-k)$ uniformly in $0 < h < h_0$ and locally uniformly in q. From (33) and (34) we conclude

$$\begin{aligned}\frac{d}{dh} Z_{q,h}(h) &= \lim_{k\to h} (h-k)^{-1}\,[Z_{q,h}(h) - Z_{q,k}(k)] \\ &= \int_0^1 Z_{q,h}(h)\,Z_{q,h}^{-1}(ht)[A(t_1+ht) + ht\dot{A}(t_1+ht) \\ &\quad + t\dot{b}(ht)\,q']\,Z_{q,h}(ht)\,dt.\end{aligned}$$

For $h \to 0$ we obtain

$$\lim_{h\to 0} \frac{d}{dh} Z_{q,h}(h) = \int_0^1 e^{(1-t)bq'}[A(t_1) + t\dot{b}(0)\,q']\,e^{tbq'}\,dt \tag{35}$$

locally uniformly in q. Since the function on the right-hand side of (35) is continuous in q, the continuity of $(d/dh)\,Z_{q,h}(h)$ is established.

9°. We construct a nonsingular $n \times n$-matrix S such that

$$X_{q,h}(\omega, 0)\cdot S = SX_{q^0,0}(\omega, 0) \tag{36}$$

for $h > 0$ sufficiently small and appropriate q. Further, we shall simply denote $X_{q,h}(\omega, 0)$ by $X_{q,h}$.

Denote s_i, $i = 1,\ldots, n$ the columns of S and choose $s_n = e_n$. Then, taking into account that by 7°

$$X_{q^0,0} = \begin{pmatrix} 0, 1,\ldots, 0 \\ \cdots\cdots\cdots \\ 0,\ldots, 1 \\ -\beta_1,\ldots, -\beta_n \end{pmatrix}$$

we see that (36) is equivalent with the set of equations

$$\begin{aligned} X_{q,h}s_1 &= -\beta_1 e_n \\ X_{q,h}s_2 &= s_1 - \beta_2 e_n \\ &\cdots\cdots\cdots\cdots \\ X_{q,h}e_n &= s_{n-1} - \beta_n e_n\,. \end{aligned}$$

or, equivalently,

$$\begin{aligned} s_{n-1} &= (X_{q,h} + \beta_n)\, e_n \\ s_{n-2} &= X_{q,h} s_{n-1} + \beta_{n-1} e_n \\ &\cdots\cdots\cdots\cdots \\ 0 &= X_{q,h} s_1 + \beta_1 e_n \,. \end{aligned} \tag{37}$$

This set of equations is equivalent with

$$s_i = \left[X_{q,h}^{n-i} + \sum_{j=0}^{n-i-1} X_{q,h}^{j} \beta_{i+j+1} \right] e_n \,, \qquad i = 1,\ldots, n-1 \tag{38}$$

$$0 = \left[X_{q,h}^{n} + \sum_{j=0}^{n-1} X_{q,h}^{j} \beta_{1+j} \right] e_n \,. \tag{39}$$

Denote $\phi(q, h) = [X_{q,h}^n + \sum_{j=0}^{n-1} \beta_{1+j} X_{q,h}^j]\, e_n$. We have $\phi(q^0, 0) = 0$, because the square bracket in (39) is the characteristic polynomial of $X_{q^0,0}$. Since $X_{q,h}$ is a continuously differentiable function of q, h, so is ϕ. Therefore, by the implicit function theorem, if we prove that $(\partial/\partial q)(\phi(q, 0)|_{q=q_0}$ is non-singular, it follows that there is a continuous function $q(h)$ for h sufficiently small such that $\phi(q(h), h) = 0$ and $q(0) = q^0$. By 2° and 7° we have

$$X_{q,0} = \begin{pmatrix} 0, 1,\ldots, 0 \\ \cdots\cdots\cdots\cdots \\ 0,\ldots, 1 \\ \alpha_1 + p_1 ,\ldots, \alpha_n + p_n \end{pmatrix}$$

where $p = (p_1 ,\ldots, p_n) = q(e^{\tilde{b}'\tilde{q}} - 1)/\tilde{b}'\tilde{q}$ and $\tilde{q} = Y(t_1)'q$ (cf. (21)). From this it follows that

$$X_{q,0}^j e_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \alpha_n + p_n \\ \alpha_{n-1} + p_{n-1} + f_{j,j+2}(p_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \alpha_{n-j+1} + p_{n-j+1} + f_{j,n}(p_n ,\ldots, p_{n-j+2}) \end{pmatrix}$$

where 1 is in the n-jth row and f_{jv} are polynomials. Consequently,

$$\phi(q, 0) = \begin{pmatrix} \alpha_n + p_n \\ \alpha_{n-1} + p_{n-1} + f_2(p_n) \\ \cdots\cdots\cdots\cdots \\ \alpha_1 + p_1 + f_n(p_n ,\ldots, p_2) \end{pmatrix} \tag{40}$$

where $f_2 ,\ldots, f_n$ are polynomials.

We have

$$\frac{\partial\phi(q,0)}{\partial q} = \frac{\partial\phi(q,0)}{\partial p}\cdot\frac{\partial p}{\partial\tilde{q}}\cdot\frac{\partial\tilde{q}}{\partial q}.$$

From (40) follows that $(\partial\phi(q,0))/\partial p$ is triangular with ones in the diagonal and, consequently, nonsingular.

Further, if we denote $\hat{p} = Y^{-1}(\omega)'p$, $\hat{q} = Y^{-1}(\omega)'\tilde{q}$, we have

$$\hat{p} = \frac{e^{(Y(\omega)\tilde{b})'\hat{q}} - 1}{(Y(\omega)\,\tilde{b})'\,\hat{q}}\cdot\hat{q} = \frac{e^{\hat{q}_n} - 1}{\hat{q}_n}\cdot\hat{q}$$

so that

$$\frac{\partial\hat{p}_i}{\partial\hat{q}_j} = \frac{e^{\hat{q}_n} - 1}{\hat{q}_n}\,\delta_{ij} \quad \text{if } j \neq n, \qquad \frac{\partial\hat{p}_n}{\partial\hat{q}_n} = \frac{\partial}{\partial\hat{q}_n}(e^{\hat{q}_n} - 1) \neq 0$$

which proves that $\partial\hat{p}/\partial\hat{q}$ is nonsingular. Consequently,

$$\partial p/\partial\tilde{q} = Y(\omega)'(\partial\hat{p}/\partial\hat{q})\,Y^{-1}(\omega)'$$

is nonsingular. Since $\partial\tilde{q}/\partial q = Y(t_1)'$ is also nonsingular, we have proved that $(\partial\phi/\partial q)(q^0, 0)$ is nonsingular.

Thus, for any q^0 we can find an $h > 0$ and q such that (39) and, consequently (36) is satisfied and q is arbitrarily close to q^0. From (37) it follows that S is a continuous function of q and h. But for $q = q^0$ and $h = 0$, $S = I$ so that for $h > 0$ sufficiently small S will be nonsingular. This completes the proof.

Remark 2. If we allow $Q(t)$ to be complex, then the characteristic multipliers can be shifted to any nonzero numbers $\sigma_1, \ldots, \sigma_n$ by closed loop control $u = Q(t)x$.

Remark 3. If $A(t)$, $B(t)$ are only continuous, the theorem is still valid in a weaker form: namely, if (1) is controllable and Σ is a spectrum such that $\sigma_1, \ldots, \sigma_n \geqslant 0$, then to any $\epsilon > 0$ there is a matrix $Q(t)$ such that the characteristic multipliers σ_i' of the system (7) satisfy $|\sigma_i' - \sigma_i| < \epsilon$. (The case (ii) of the theorem can be changed in a similar manner). Also the sufficiency part remains valid.

This can be seen from the fact that (7) can still be proved in a weaker form

$$X_{Q,h}(\omega, 0) = Y(\omega)\cdot\prod_{i=1}^{r} e^{\tilde{b}_i\tilde{q}_i'} + \theta(h) = \psi_{Q,0}(\omega, 0) + \theta(h) \qquad (41)$$

where $\lim_{h\to 0} \theta(h) = 0$ locally uniformly in q. Thus, the steps 2°–7° of the proof can be repeated without change and we can find the Q^0 such that $X_{Q^0,0}(\omega, 0)$ has spectrum Σ, or an arbitrary close spectrum to Σ, if Σ contains zero elements. Since the spectrum of a matrix is a continuous function of its entries the statements follow from (41).

Remark 4. The matrix, which we have constructed to solve our problem, is discontinuous. This is not essential and it can be verified that there is a C^∞-matrix $Q(t)$ which solves the problem. For this purpose, let us first note that the function $\xi_\delta(t)$ of 5° can be chosen C^∞ with all required properties preserved. Choosing such a function the proof can be carried out essentially in the same way (with some calculations, of course, more complicated) with $Q(t)$ replaced by

$$\hat{Q}_h(t) = \begin{cases} h^{-1}Q^{(j)}\xi_\delta[h^{-1}(t - t_j)] & \text{in } [t_j + \nu\omega, t_j + \nu\omega + h], \nu \text{ integer} \\ 0 & \text{elsewhere.} \end{cases}$$

References

1. Kalman, R. E., Ho, Y. C., and Narendra, K. S., Controllability of linear dynamical systems. Contr. Diff. Eqs. **1** (1963), 189–213.
2. Popov, V. M., Hyperstability and optimality of automatic systems with several control functions. Rev. Roumaine Sci. Techn., Electrotechn. et Energ., **9** (1964), pp.629–690.
3. Popov, V. M., Hyperstabilitatea sistemelor automate. Editura Academiei Republici Socialiste Romania, Bucuresti 1966.
4. Wonham, W. M., On pole assignment in multi-input controllable linear systems, Brown University. Division of Applied Mathematics, Technical Report 67-2, (1967).
5. Langenhop, C. E., On the stabilization of linear systems. *Proc. Am. Math. Soc.* **15** (1964), pp. 735–742.
6. Brunovsky, P., On stabilization of linear systems under a certain class of persistent perturbations. *Differentsialnyie uravnenia* **2** (1966), 769–777.
7. Lee, E. B., Markus, L., "Foundations of Optimal Control Theory." Wiley, New York, 1967.

P. Brunovský

A Classification of Linear Controllable Systems

Kybernetika 6(3) (1970), 173–188.

KYBERNETIKA ČÍSLO 3, ROČNÍK 6/1970

A Classification of Linear Controllable Systems*

PAVOL BRUNOVSKÝ

The concept of feedback (*F*-) equivalence of linear controllable systems is defined and the classification of such systems based upon this equivalence concept is discussed.

1. INTRODUCTION

This paper is concerned with linear control systems, which can be represented by systems of linear differential equations of the form

$$\dot{x} = Ax + Bu \tag{1}$$

where x and u are n- and m-vectors respectively, A and B are matrices of appropriate size, in general time-dependent.

Since the system (1) is uniquely determined by the pair of matrices A, B, we shall frequently call it $\langle A, B\rangle$.

The basic question, which leads to the classification studied in this paper, can be formulated as follows:

Having two systems $\langle A, B\rangle$ and $\langle A', B'\rangle$, is there a linear feedback which, added to $\langle A, B\rangle$, yields a system, which behaves like $\langle A', B'\rangle$?

We shall make this question more precise in the next section. At this point let us note that the feedback will be required to be constant, time varying or periodic in t according to the system itself.

We shall see, that this question gives rise to a classification of controllable autonomous systems into a finite number of classe seach of which can be represented by a very simple canonical form. For time-varying systems, such a classification will be given for an important subclass of controllable systems.

* This research was supported by NASA under grant No. NGR-24-005-063 during the author's stay at Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota, U. S. A.

174 Let us note that this paper is related to [1], [2], [3], where similar concepts of equivalence have been introduced, though for different purposes. For time-varying systems, this paper extends a result of [4]. The point of view of studying control system is to a certain extent related to that of [5].

2. AUTONOMOUS SYSTEMS

In this section, we shall assume that A, B are constant matrices and that the system is completely controllable, i. e.,

$$\text{rank}\,(B, AB, \ldots, A^{n-1}B) = n\,. \tag{2}$$

To formulate the question raised in the preceding section more precisely, we translate it into an algebraic form.

By adding a linear feedback to $\langle A, B\rangle$, we mean that in (1), we substitute $u = Qx + v$, where Q is $m \times n$ constant. As a result of this transformation, we obtain a system $\langle A'', B''\rangle$, with $A'' = A + BQ$, $B'' = B$.

By saying that $\langle A'', B''\rangle$ behaves like $\langle A', B'\rangle$ we mean that by nonsingular linear transformations of the state (output) and input variables, $\langle A'', B''\rangle$ can be brought into $\langle A', B'\rangle$ or, algebraically, there are nonsingular matrices C and D of type $n \times n$, $m \times m$ respectively, such that $A' = C^{-1}A''C$, $B' = C^{-1}BD$.

Summarizing, we find that the question of the preceding section asks, whether for given systems $\langle A, B\rangle$, $\langle A', B'\rangle$ there are matrices $C(m \times n)$, $Q(m \times n)$, $D(m \times m)$, C, D being nonsingular, such that

$$A' = C^{-1}(A + BQ)\,C\,, \quad B' = C^{-1}BD\,. \tag{3}$$

If the answer is positive, we shall say that $\langle A, B\rangle$ and $\langle A', B'\rangle$ are feedback (or, briefly, F-) equivalent.

By a straightforward computation it can be checked that F-equivalence is actually an equivalence relation, i. e., it is symmetric, reflexive and transitive. Moreover, the order of transformations, by which $\langle A', B'\rangle$ is obtained from $\langle A, B\rangle$, can be changed.

It will be frequently convenient to express that a system $\langle A', B'\rangle$ can be obtained from $\langle A, B\rangle$ by one of the transformations, occuring in the definition of F-equivalence, only.

Referring to our definition, $\langle A, B\rangle$ and $\langle A', B'\rangle$ will be called

– C-equivalent, if $Q = 0$, $D = E$,

– Q-equivalent, if $C = E_n$, $D = E_m$,

– D-equivalent, if $C = E$, $Q = 0$,

where E is the unity matrix, the index indicating its size. It is easy to check, that these relations are equivalence relations.

We associate with a given system $\langle A, B\rangle$ n numbers $r_0, r_1, \ldots, r_{n-1}$ as follows: 175

$$(4)\quad r_0 = \operatorname{rank} B\,,$$
$$r_j = \operatorname{rank}(B, AB, \ldots, A^jB) - \operatorname{rank}(B, AB, \ldots, A^{j-1}B)\,,\quad 1 \leqq j \leqq n-1\,.$$

Geometrically, if we denote by $L_j\langle A, B\rangle$ the linear subspace of R^n, spanned by the column vectors of $B, AB, \ldots, A^jB$, by Λ_j the orthogonal complement of L_{j-1} in L_j and by $\pi_j(f)$ the orthogonal projection of a vector f into Λ_j, then r_j is the dimension of Λ_j, which is equal to rank $(\pi_j(A^jB))$.

Obviously, $0 \leqq r_j \leqq m$ for $0 \leqq j \leqq n-1$ and, because of (2), $\sum_{j=0}^{n-1} r_j = n$. Moreover, since $A^jb_i = \sum_{\nu\in I} A^jb_\nu$, $I \in \{1, \ldots, m\}$ implies $A^{j+1}b_i = \sum_{\nu\in I} A^{j+1}b_\nu$, we have $r_0 \geqq r_1 \geqq, \ldots, \geqq r_{n-1}$ and we can choose a basis S of R^n from the column vectors of $(B, AB, \ldots, A^{n-1}B)$ in such a way that the vectors $\{\pi_j(A^jb_i) | A^jb_i \in S,\ j \text{ fixed}\}$ span Λ_j (consequently, their number is r_j) and if $A^jb_i \notin S$, than $A^{j+1}b_i \notin S$. Such a basis we shall call pyramidal.

Assuming that we have chosen a pyramidal basis S, we can associate with every column b_i a number p_i, such that $A^jb_i \in S$ for $0 \leqq j \leqq p_i - 1$, but $A^{p_i}b_i \notin S$. By re-ordering suitably the columns of B (this is a D-transformation) we can achieve that $p_1 \geqq p_2 \geqq, \ldots, \geqq p_m$. Consequently, the p-numbers can be uniquely determined by the r-numbers, associated with $\langle A, B\rangle$, as follows:

$$(5)\qquad p_i \text{ is the number of } r_j\text{'s, which are } \geqq i\,.$$

Conversely, the r-numbers are evidently uniquely determined by the p-numbers of $\langle A, B\rangle$.

As a result of our discussion we have:

Lemma 1. *For every controllable system $\langle A, B\rangle$, the finite sequences of numbers $R\langle A, B\rangle = \{r_j\}_{j=0}^{n-1}$, $P\langle A, B\rangle = \{p_i\}_{i=1}^{m}$, defined respectively by* (4), (5), *have the following properties*:

(i) $0 \leqq r_j \leqq m$, $r_0 \geqq r_1 \geqq \ldots \geqq r_{p_1-1} > 0$, $r_j = 0$ *for* $j \geqq p_1$, $\sum_{j=0}^{n-1} r_j = n$,

(ii) $0 \leqq p_i \leqq n$, $p_1 \geqq \ldots \geqq p_{r_0} > 0$, $p_i = 0$ *for* $i > p_{r_0}$, $\sum_{i=1}^{m} p_i = n$,

(iii) $P\langle A, B\rangle = P\langle A', B'\rangle$ *if and only if* $R\langle A, B\rangle = R\langle A', B'\rangle$.

(iv) *There is a system $\langle A, B'\rangle$ D-equivalent with $\langle A, B\rangle$ such that the vectors A^jb_i, $1 \leqq i \leqq r_0$, $0 \leqq j \leqq p_i - 1$ form a basis of R_n.*

Now we are able to formulate

176 **Theorem 1.** *$\langle A, B\rangle$ is F-equivalent with $\langle A', B'\rangle$ if and only if $R\langle A, B\rangle = R\langle A', B'\rangle$ (or, $P\langle A, B\rangle = P\langle A', B'\rangle$).*

Theorem 2. *Let $P\langle A, B\rangle = \{p_i\}_{i=1}^{m}$, $R\langle A, B\rangle = \{r_j\}_{j=0}^{n-1}$. Then, $\langle A, B\rangle$ is F-equivalent with a decoupled system of r_0 integrators:*

$$(6)\qquad \dot{y}_{k_i+1} = y_{k_i+2}, \ldots, \dot{y}_{k_{i+1}-1} = y_{k_{i+1}}, \quad \dot{y}_{k_{i+1}} = v_{i+1}, \quad i = 0, \ldots, r_0 - 1$$

where $k_i = \sum_{\nu=1}^{i} p_\nu$.

We first prove theorem 2, with the aid of

Lemma 2. *Let $P\langle A, B\rangle = \{p_i\}_{i=1}^{m}$, $R\langle A, B\rangle = \{r_j\}_{j=0}^{n-1}$. Then, $\langle A, B\rangle$ is C-equivalent with a system $\langle A', B'\rangle$ of the following form:*

$$A' = \begin{pmatrix} A'_{11}, \ldots, A'_{1r_0} \\ \ldots\ldots\ldots \\ A'_{r_0 1}, \ldots, A'_{r_0 r_0} \end{pmatrix}, \quad B' = \begin{pmatrix} B'_1 \\ \vdots \\ B'_{r_0} \end{pmatrix}$$

where A'_{ij} are $p_i \times p_j$,

$$A'_{ij} = \begin{pmatrix} 0 & \ldots, 0 \\ \ldots\ldots\ldots\ldots \\ 0 & \ldots, 0 \\ \alpha_{ik_{i-1}+1}, & \ldots, \alpha_{ik_j} \end{pmatrix}$$

if $i \neq j$,

$$A'_{ii} = \begin{pmatrix} 0 & , 1 & \ldots, 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0 & , 0 & \ldots, 1 \\ \alpha_{ik_{i-1}+1}, & \alpha_{ik_{i-1}+2}, & \ldots, \alpha_{ik_i} \end{pmatrix},$$

$k_i = \sum_{\nu=1}^{i} p_\nu$, *$B'_i$ are $p_i \times m$,*

$$B'_i = \begin{pmatrix} 0, \ldots\ldots\ldots\ldots, 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0, \ldots\ldots\ldots\ldots, 0 \\ 0, \ldots, 0, 1, \gamma_{ii+1}, \ldots, \gamma_{im} \end{pmatrix}$$

(1 is in the i-th column), $i, j = 1, \ldots, m$.

This lemma is proved in [6] and, in fact, is also a special case of lemma 8 of this paper.

Proof of theorem 2. In virtue of lemma 1, we can assume that $\langle A, B\rangle$ has the special form, given in its formulation. Denote by $\tilde{B}$ the submatrix of B, consisting of its k_i-th rows, $i = 1, \ldots, r_0$ (those are precisely the non-zero rows of B), by $\tilde{\tilde{B}}$ the

submatrix of $\tilde{B}$, consisting of its first r_0 columns. $\tilde{\tilde{B}}$ is a nonsingular triangular $r_0 \times r_0$ matrix and, furthermore, the columns $\tilde{b}_i$, $i > r_0$ of $\tilde{B}$ are linear combinations of the columns of $\tilde{\tilde{B}}$, i. e., there are r_0-vectors d_i, such that $-\tilde{b}_i = \tilde{\tilde{B}}\tilde{d}_i$, $r_0 < i \leqq m$. Denote

$$D = \begin{pmatrix} \tilde{\tilde{B}}^{-1}, \tilde{d}_{r_0+1}, \ldots, \tilde{d}_m \\ 0 \quad , E_{m-r_0} \end{pmatrix}.$$

Then, $\tilde{B}D = (E_{r_0}, 0)$ and, consequently, $B' = BD$ has all elements zero except for $b'_{k_i i}$, which are equal 1; $\langle A, B' \rangle$ is D-equivalent with $\langle A, B \rangle$. To complete the proof, we define Q as follows: $q_{ij} = -\alpha_{ij}$, $1 \leqq i \leqq r_0$, $1 \leqq j \leqq n$, $q_{ij} = 0$, $r_0 < i \leqq m$, $1 \leqq j \leqq n$. Then, the system $\langle A', B' \rangle$ with $A' = A + B'Q$ is F-equivalent with $\langle A, B \rangle$ and has the required form.

Proof of theorem 1. The F-equivalence of the systems with the same R (or P) follows directly from theorem 2.

To prove the opposite implication, it suffices to show that none of the C-, D-, Q-transformations changes the r-numbers of a controllable system.

For C- and D-transformations, this statement is trivial. If $\langle A', B \rangle$ is a Q-transform of $\langle A, B \rangle$, $A' = A + BQ$, then we have $A'^j B = (A + BQ)^j B = A^j B + G$, where G is an $n \times m$ matrix, whose columns are contained in $L_{j-1}\langle A, B \rangle$ and the statement follows by induction in j.

The system (6) of theorem 2 can be considered as a canonical form for a particular class of systems.

Let us also note that besides other things, theorem *1* justifies our restriction to controllable systems. Namely, it shows that there is no C-, D- or Q-transformation, which will make a controllable system from a non-controllable one and conversely.

It is apparent that for given m and n, there is only a finite number of equivalence classes. From lemma 1 and theorem 1 it follows that the number of classes is equal to the number ϱ_{mn} of ways in which n can be written as a sum of n nonnegative integers (or the number of ways, in which n can be written as a sum of n nonnegative integers, not exceeding m). However, there is no formula known for the computation of ϱ_{mn}. The problem of finding the numbers ϱ_{mn} is an old numbertheoretical problem called "partition problem". It goes back to Euler, who gave a "generating function" for ϱ_{mn}, which is

$$F_m(x) = \frac{1}{(1-x)(1-x^2)\ldots(1-x^m)}.$$

This means that if we expand F formally into Taylor series, $F_m(x) = 1 + \sum_{n=1}^{\infty} \varrho_{mn} x^n$, than ϱ_{mn} are the numbers, defined above. For details, cf. [10].

The following two corollaries illustrate the usefulness of theorems 1 and 2. For other applications of theorem 2 (or, rather, lemma 2) the reader is referred to [6], [7].

178 **Corollary** *1* (cf. [1], [2], [7], [8], [9]). *To any n-th degree polynomial* $P(\lambda)$ *and any controllable system* $\langle A, B\rangle$, *there is a system* $\langle A', B\rangle$, *Q-equivalent with* $\langle A, B\rangle$, *such that the characteristic polynomial of* A' *is* $P(\lambda)$. *In particular, every controllable system can be stabilized by an appropriate linear feedback.*

Let us note, that also the converse is true, but we shall not prove it here (cf. [1], [2], [9]).

Proof. Since C- and D-transformations do not change the characteristic polynomial of A, we can assume that $\langle A, B\rangle$ is in the canonical form (6). Then, we define the Q-transformation by $v_i = x_{k_i+1} + w_i$, $1 \leqq i \leqq r_0 - 1$, $v_{r_0} = -(\beta_1 x_1 + \ldots + \beta_n x_n) + w_{r_0}$, where $P(\lambda) = \lambda^n + \beta_1 \lambda^{n-1} + \ldots + \beta_{n-1}$.

Corollary 2 (cf. [7]). *If* $\langle A, B\rangle$ *is controllable, then for any non-zero vector* $f \in L_0\langle A, B\rangle$, *there is an* $m \times n$*-matrix* Q *such that* $\langle A + BQ, f\rangle$ *is controllable.*

Proof. We can obviously again assume that $\langle A, B\rangle$ is in the canonical form (6). Let $f = \sum_{i=1}^{r_0} \lambda_i b_i$, and let $\lambda_l \neq 0$. Then, if $l \neq r_0$, we define

$$q_{ij} = \begin{cases} 1 \text{ if either } i \neq l, i \neq r_0, j = k_i + 1, \text{ or } i = r_0, j = 1, \\ 0 \text{ otherwise.} \end{cases}$$

If $l = r_0$, we define

$$q_{ij} = \begin{cases} 1 \text{ if } j = k_i + 1, i \neq r_0, \\ 0 \text{ otherwise.} \end{cases}$$

It can be easily verified that the matrix $H = (f, A'f, \ldots, A'^{n-1}f)$, where $A' = A + BQ$, has the following form:

$$H = \begin{pmatrix} H_{11}, 0 \\ H_{21}, H_{22} \end{pmatrix}$$

where H_{11} is $k_l \times k_l$, both H_{11} and H_{22} are supper triangular with λ_l on the diagonal. Consequently, H is non-singular, q. e. d.

3. TIME-VARYING SYSTEMS

In this section, we apply some of the ideas of the preceding section to time-varying systems.

We consider systems

$$\dot{x} = A(t)\, x + B(t)\, u \tag{7}$$

where $A(t)$ and $B(t)$ are defined and of continuity class $\mathscr{C}^\infty$ on some interval J.

For time-varying systems, we modify the concept of F-equivalence, by allowing the matrices C, Q, D to be time-varying. So, (7) and 179

$$\dot{y} = A'(t)\, y + B'(t)\, v \tag{8}$$

will be called F-equivalent on J, if there are matrices $C(t)$, $Q(t)$ and $D(t)$ on J, all of class $\mathscr{C}^\infty$, such that the transformations $u = Qx + Dv$ and $x = Cy$ bring (7) into (8). Translated completely into the language of time-varying matrices, defining the system, $\langle A, B\rangle$ will be said to be F-equivalent with $\langle A', B'\rangle$ on J, if there are $\mathscr{C}^\infty$-matrices C, Q, D on J such that $A' = C^{-1}[(A + BQ) - \dot{C}]\, C$, $B' = C^{-1}BD$ for $t \in J$.

It can again be easily verified that F-equivalence is actually an equivalence relation. In a similar way, the definitions of C-, Q-, D-equivalence can be modified.

There is no reason to expect that a classification of controllable time varying systems similar to that of autonomous ones can be obtained, whatever geometric definition of controllability we use. This is partly due to the fact, that for none of those concepts of controllability there is an equivalent algebraic condition, corresponding to (2).

However, there is a condition for time-varying systems, generalizing (2), which implies controllability in any reasonable geometric sense. This condition can be formulated as follows:

For F being a $\mathscr{C}^\infty$ $n \times s$-matrix function for some s, denote $\mathfrak{A}\, F(t) = A(t)\,(Ft) + \dot{F}(t)$. Then, for every t,

$$\operatorname{rank}\,(B, \mathfrak{A}B, \ldots, \mathfrak{A}^{n-1}B) = n\,. \tag{9}$$

Let us now ask the following question: When is a time-varying system F-equivalent to an autonomous one? Those systems, which are equivalent to autonomous ones, we are of course able to classify. It turns out, that the systems, F-equivalent to autonomous ones, are precisely those, which satisfy a somewhat strengthened condition (9).

For time-varying systems, we define $R\langle A, B\rangle$ as the n-tuple of functions $\{r_j(t)\}_{j=0}^{n-1}$ on J, where

$$r_j(t) = \operatorname{rank}\,(B(t), \ldots, \mathfrak{A}^j B(t)) - \operatorname{rank}\,(B(t), \ldots, \mathfrak{A}^{j-1} B(t))\,.$$

$P\langle A, B\rangle$ is the m-tuple of functions $p_i(t)$ defined for every t by (5). L_j, Λ_j and π_j are defined similarly as for autonomous systems with A replaced by $\mathfrak{A}$. They also depend on t.

Theorem 3. *The system $\langle A, B\rangle$ is F-equivalent with a controllable autonomous system on J if and only if the functions $r_0(t), \ldots, r_{n-1}(t)$ from $R\langle A, B\rangle$ are constant on J and $r_0(t) + \ldots + r_{n-1}(t) = n$.*

For better orientation, we divide the proof of this theorem into several lemmas.

180 **Lemma 3.** *Denote* $S_j(t) = \{\mathfrak{A}^\nu b_i(t) | 0 \leqq \nu \leqq j, i \in M_\nu\}$, *where* M_ν *are subsets of* $\{1, \ldots, m\}$. *Let* $S_j(t_0)$ *be a basis of* $L_j(t_0)$. *Then,*

(i) $S_j(t)$ *is a basis of* $L_j(t)$ *in a neighbourhood* U *of* t_0.

(ii) *If* $f(t) \in L_j(t)$ *and* f *is* $\mathscr{C}^\infty$ *on* U, *then*

$$f(t) = \sum_{\substack{0 \leqq \nu \leqq j \\ i \in M_\nu}} \gamma_{i\nu} \mathfrak{A}^\nu b_i(t)$$

where $\gamma_{i\nu}$ *are* $\mathscr{C}^\infty$ *functions on* U *and* $b_i(t)$ *are the columns of* $B(t)$.

Proof. Let $\varrho_j = \sum_{\nu \leqq j} r_\nu$. Then, there is a $\varrho_j \times \varrho_j$ nonsingular submatrix $\tilde{S}_j(t_0)$ of $S_j(t_0)$. Since $\det \tilde{S}_j(t)$ is continuous, we have $\det \tilde{S}_j(t_0) \neq 0$ in some neighbourhood U of t_0. Since $\gamma_{i\nu}$ can be expressed as polynomials of the entries of $f(t)$ and $S_j(t)$ divided by $\det \tilde{S}_j(t)$, the lemma follows.

Chose a norm $\|.\|$ in R^n. As an immediate consequence of lemma 3 we obtain

Lemma 4. *Let the functions of* $R\langle A, B\rangle$ *be constant on* J. *Then, the subspaces* $L_j(t)$ *and* $\Lambda_j(t)$ *depend continuously on* t *on* J *in the following sense:*

To any ball $N_R = \{x | \|x\| \leqq R\}$ *in* R^n, *any* $t_0 \in J$ *and* $\varepsilon > 0$ *there is a* $\delta > 0$ *such that for* $|t - t_0| < \delta$,

$$L_j(t) \cap N_R, \quad L_j(t_0) \cap N_R, \quad \Lambda_j(t) \cap N_R, \quad \Lambda_j(t_0) \cap N_R$$

are contained in the ε*-neighbourhoods of* $L_j(t_0)$, $L_j(t)$, $\Lambda_j(t_0)$, $\Lambda_j(t)$ *respectively.*

Lemma 5. *If* $\langle A, B\rangle$ *and* $\langle A, B'\rangle$ *are D-equivalent, then* $L_j\langle A, B\rangle(t) = L_j\langle A, B'\rangle(t)$ *and* $\Lambda_j\langle A, B\rangle(t) = \Lambda_j\langle A, B'\rangle(t)$ *for all* $t \in J$.

Proof. We have $\mathfrak{A}^j b_i' = \sum_{\mu=1}^m \mathfrak{A}^j(b_\mu d_{\mu i}) = \sum_{\mu=1}^m (\mathfrak{A}^j b_\mu)\, d_{\mu i} - \sum_{\mu=1}^m (\mathfrak{A}^{j-1} b_\mu)\, \dot{d}_{\mu i}$. From this we obtain by induction $L_j\langle A, B'\rangle \subset L_j\langle A, B\rangle$, $0 \leqq j \leqq n-1$. By symmetry of D-equivalence, we have the opposite inclusion and, thus, equality. The equality of Λ_j's follows trivially.

Lemma 6. *Let for some* t, $S_j(t)$ *be a pyramidal basis of* $L_j(t)$. *Then,* $S_j(t)$ *can be completed into a pyramidal basis of* $L_{n-1}(t)$.

Proof. We prove that $S_j(t)$ can be extended into a pyramidal basis $S_{j+1}(t)$ of $L_{j+1}(t)$. The rest follows then by induction.

Let $\mathfrak{A}^j b_i \in S_j(t)$, if and only if $i \in M_j$. Then we have for every b_k, $1 \leqq k \leqq m$, $\mathfrak{A}^j b_k = \sum_{i \in M_j} \lambda_i \mathfrak{A}^j b_i + g_k$, where $g_k \in L_{j-1}(t)$.

Thus,

$$\mathfrak{A}^{j+1} b_k = \sum_{i \in M_j} \lambda_i \mathfrak{A}^{j+1} b_i - \sum_{i \in M_j} \dot{\lambda}_i \mathfrak{A}^j b_i + \mathfrak{A} g_k = \sum_{i \in M_j} \lambda_i \mathfrak{A}^{j+1} b_i + f_k$$

where $f_k \in L_j(t)$. From this it is clear that to complete the basis for $S_{j+1}(t)$, we can add to $S_j(t)$ any $r_{k+1}(t)$ linearly independent vectors from $\mathfrak{A}^{j+1}b_i$, $i \in M_j$, q. e. d. 181

Corollary 3. $m \geqq r_0(t) \geqq r_1(t) \geqq \ldots \geqq r_{n-1}(t) \geqq 0$ *for all* t.

Lemma 7. *Let the functions of* $R\langle A, B\rangle$ *be constant on* J. *Then,* $\langle A, B\rangle$ *is D-equivalent with a system* $\langle A, B'\rangle$ *such that*

$S_i'(t) = \{\mathfrak{A}^v b_i'(t) | 0 \leqq v \leqq j, \quad 1 \leqq i \leqq r_v\}$ *are bases of* $L_j(t)$ *for* $t \in J$.

Note that the theorem of [11] is a special case of lemma 7.

Proof. The matrix $D(t)$ will be constructed as a product of n matrices $D = D_0 \ldots$ $\ldots D_{n-1}$ in such a way that for $\langle A, BD_0 \ldots . D_k\rangle$ the statement of the lemma will be valid for $j \leqq k$.

Assume that we have already constructed the matrices $D_0, \ldots, D_{k-1}$. In virtue of lemmas 3, 6 we can cover J by a sequence of open intervals $J_\mu = (a_\mu, b_\mu)$, $-\infty < \mu < \infty$ in such a way that $\frac{1}{2}(a_\mu + b_\mu)$ is an increasing sequence, $b_\mu < a_{\mu+2}$ for all μ and for every μ there is a subset $M_\mu \subset \{1, \ldots, r_{k-1}\}$ such that $\{\pi_k(\mathfrak{A}^k b_i(t)) | i \in M_\mu\}$ span $\Lambda_k(t)$ for $t \in J_\mu$. D_k will again be constructed in steps. First we define $D_k(t)$ on $[b_{-1}, a_1]$ as follows.

We put $d_i = e_{\sigma i}$, $1 \leqq i \leqq r_k$, $d_i = e_i$ for $r_{k-1} < i \leqq m$ and the remaining d_i's we put equal to the remaining e_i's arbitrarily. There $M_0 = \{\sigma_1, \ldots, \sigma_{r_k}\}$, d_i, e_i are the columns of D_k, E_m respectively. By multiplication of B by D_k, the columns $\{b_i | i \in M_0\}$ are brought into the first r_k positions and the columns b_i, $i > r_{k-1}$ remain without change.

We proceed by introduction assuming that $D_k(t)$ has been constructed on $[b_{-\mu}, a_\mu]$ with following properties: (which are obviously satisfied for $\mu = 1$).

Denote $\tilde{D}_\mu(t) = D_k(t)$ for $t \in [b_{-\mu}, a_\mu]$, $\tilde{D}_\mu(t) = D_k(b_{-\mu})$ for $t < b_{-\mu}$, $\tilde{D}_\mu(t) = D_k(a_\mu)$ for $t > a_\mu$, $\tilde{B}(t) = B(t)\,\tilde{D}_\mu(t)$. Then:

(i) $D_\mu(t)$ is nonsingular and $\mathscr{C}^\infty$ for all t

(ii) The vectors $\{\mathfrak{A}^v \tilde{b}_i | 0 \leqq v \leqq k-1, 1 \leqq i \leqq r_v\}$ span L_{k-1} for all t

(iii) The vectors $\{\pi_k(\mathfrak{A}^j b_i) | 1 \leqq i \leqq r_k\}$ span $\Lambda_k(t)$ for $a_{-\mu+1} \leqq t \leqq b_{\mu-1}$ and there are subsets M_μ, $M_{-\mu}$ of $\{1, \ldots, r_{k-1}\}$ such that

$$\{\pi_k(\mathfrak{A}^k \tilde{b}_i) | i \in M_\mu\}, \ \{\pi_k(\mathfrak{A}^k \tilde{b}_i) | i \in N_\mu\} \text{ span } \Lambda_k(t) \text{ for } t \in [a_\mu, b_\mu]$$

and $t \in [a_{-\mu}, b_{-\mu}]$ respectively.

We show that $D_k(t)$ can be extended so the interval $[b_{-\mu+1}, a_{\mu+1}]$ in such a way that (i)–(iii) remains valid with μ replaced by $\mu + 1$. We show only the extension forwards, the extension backwards being entirely similar.

We divide $[a_\mu, b_{\mu-1}]$ into r_k subintervals of equal length $\tau = r_j^{-1}(b_{\mu-1} - a_\mu)$ and denote $\alpha_\zeta = a_\mu + \zeta\tau$. Again $D_k(t)$ will be extended by induction. We assume

182 that $D_k(t)$ has been extended for $t \leqq a_{\zeta-1}$, $\zeta \leqq n$ in such a way that if we denote

$$\bar{D}_{\zeta-1}(t) = D_k(t) \text{ for } t \leqq \alpha_{\zeta-1},\ \bar{D}_{\zeta-1}(t) = D_k(\alpha_{\zeta-1}) \text{ for } t \geqq \alpha_{\zeta-1},$$

$\bar{B}(t) = B(t)\,\bar{D}_{\zeta-1}(t)$, then we have

(a) $\bar{D}_{\zeta-1}(t)$ is $\mathscr{C}^\infty$, nonsingular for all t,

(b) The vectors $\{\mathfrak{A}^\nu \bar{b}_i | 0 \leqq \nu \leqq k-1,\ 1 \leqq i \leqq r_\nu\}$ span $L_{k-1}(t)$ for all t,

(c) The vectors $\{\pi_k(\mathfrak{A}^k \bar{b}_i) | 1 \leqq i \leqq r_k\}$ span $\Lambda_k(t)$ for $t \leqq b_{\mu-1}$,

(d) There is a subset N_ζ of $(\zeta, \ldots, r_{k-1}\}$ such that

$\{\pi_j(\mathfrak{A}^j \bar{b}_i) | 1 \leqq i < \zeta \text{ or } i \in N_\zeta\}$ is a basis of $\Lambda_k(t)$ for $\alpha_{\zeta-1} \leqq t \leqq b_\mu$.

We show that $D_k(t)$ can be extended for $t \leqq \alpha_\zeta$ so that (a)–(d) remain valid with $\zeta - 1$ replaced by ζ.

From (c) and (d) it follows that there is an $l \in N_\zeta$ such that $\mathfrak{A}^j \bar{b}_l$ is not contained in the subspace $\Delta(t)$ of $\Lambda_k(t)$, spanned by $\{\pi_k(\mathfrak{A}^k \bar{b}_i) | 1 \leqq i \leqq r_k,\ i \neq \zeta\}$ for any $t \in [\alpha_{\zeta-1}, b_{\mu-1}]$. If $l = \zeta$, we simply define $D_k(t) = \bar{D}_{\zeta-1}$ for $t \in [\alpha_{\zeta-1}, \alpha_\zeta]$, i. e. we extend $D_k(t)$ continuously as a constant. If $l \neq \zeta$ (note that then $l > r_k$) then $\pi_k(\mathfrak{A}^k \bar{b}_\zeta)$ and either $\pi_k(\mathfrak{A}^k \bar{b}_l)$ or its negative lie in the interior of the same one of the two halfspaces, into which $\Delta(t)$ divides $\Lambda_k(t)$, for all $t \in [\alpha_{\zeta-1}, \alpha_\zeta]$. In the first case, we define for $t \in [\alpha_{\zeta-1}, \alpha_\zeta]$, $D_k(t) = \bar{D}(t)\,Z(t)$, where $Z(t) = (z_{ij}(t))$ is defined as follows:

$$z_{\zeta\zeta}(t) = 1 - \psi(\alpha_{\zeta-1} + t),\quad z_{\zeta l} = -\psi(\alpha_{\zeta-1} + t), \tag{10}$$

$$z_{l\zeta}(t) = \psi(\alpha_{\zeta-1} + t),\ z_{ll} = 1 - \psi(\alpha_{\zeta-1} + t),$$

$z_{ij}(t) = \delta_{ij}$ otherwise and $\psi(t)$ is a nonnegative $\mathscr{C}^\infty$ real function such that $\psi(0) = 0$ for $t \leqq 0$, $\psi(t) = 1$ for $t \geqq \tau$, $0 \leqq \psi(t) \leqq 1$ for $0 \leqq t \leqq \tau$. In the second case, we define $z_{\zeta l}$ and $z_{l\zeta}$ with opposite signs.

Now, the validity of (a) for ζ is obvious. For the proof of (b)–(d) denote $\bar{\bar{B}}(t) = B(t)\,\bar{\bar{D}}(t)$, where $\bar{\bar{D}}$ is defined as $\bar{D}(t)$ with $\zeta - 1$ replaced by ζ. We have

$$\mathfrak{A}^\nu \bar{\bar{B}} = \mathfrak{A}^\nu(\bar{B}Z) = (\mathfrak{A}^\nu \bar{B})\,Z + F$$

where $F \in L_{\nu-1}(t)$, which implies

$$\pi_\nu(\mathfrak{A}^\nu \bar{\bar{B}}) = (\pi_\nu(\mathfrak{A}^\nu \bar{B}))\,Z\,. \tag{11}$$

If we denote $\hat{\bar{B}}$, $\hat{\bar{\bar{B}}}$ the submatrices of $\bar{B}$, $\bar{\bar{B}}$ respectively, formed by their first r_ν columns and $\hat{Z}_\nu$ the submatrix of Z, formed by its first r_ν columns and r_ν rows then, because all elements of the first r_ν columns, not belonging to $\hat{Z}_\nu$, are zero, we have from (11)

$$\pi_\nu(\mathfrak{A}^\nu \hat{\bar{\bar{B}}}_\nu) = (\pi_\nu(\mathfrak{A}^\nu \hat{\bar{B}}_\nu))\,Z_\nu \text{ for } 0 \leqq \nu \leqq k-1$$

which implies (b).

(c) follows from the fact that by (10) and (11), $\pi_k(\mathfrak{A}^k \bar{b}_\zeta)$ is for all t a convex combination of two vectors, both of which lie in the interior of the same one of the half-spaces, into which $\Delta(t)$ divides $A_k(t)$ and the other vectors of $\{\pi_k(\mathfrak{A}^k \bar{b}_i) | 1 \leqq i \leqq r_k\}$ are not affected by multiplication of $\bar{B}$ by Z. 183

(d) follows from the fact that for $t \geqq a_\zeta$, $\pi_k(\mathfrak{A}^k \bar{b}_i) = \pi_k(\mathfrak{A}^k \bar{b}_\zeta)$.

After extending $D_k(t)$ for $t \leqq b_{\mu-1}$ stepwise in the described way, we extend $D_k(t)$ for $t \leqq a_{\mu+1}$ by setting $D_k(t) = D_k(b_{\mu-1})$ for $t \in [b_{\mu-1}, a_{\mu+1}]$. Then, it is easy to verify that D_k satisfies (i)–(iii) for $t \leqq a_{\mu+1}$.

This, by introduction, proves the existence of $D_k(t)$ and, thus, of $D(t)$.

Remark 1. By a refinement of the above argument it can be shown that if the matrix $(B, \mathfrak{A}B, \ldots, \mathfrak{A}^{n-1}B)$ and all its p derivatives $(p = \max_{1 \leqq i \leqq m} p_i)$ are bounded on J and for each t there is an $n \times n$-subdeterminant of this matrix, the absolute value of which is bounded below by a positive constant, independent of t on J, then $D(t)$ can be constructed in such a way that $(B', \mathfrak{A}B', \ldots, \mathfrak{A}^{n-1}B')$ is bounded and $|\det S_{n-1}(t)|$ is bounded below by a positive constant, independent of t. This will be seen to be important for the stabilization problem.

The refinement is essentially based on the facts, that under the above boundedness assumptions the covering $\{J_\mu\}$ of J, which occurs in the proof of Lemma 7 can be for every k constructed in such a way that on every interval J_μ the absolute value of some pyramidal basis and also the length of the intersections of the consecutive intervals are bounded below by a positive constant, which is independent on μ.

Lemma 8. *Let for all $t \in J$, the vectors $\{\mathfrak{A}^\nu b_i | 0 \leqq \nu \leqq n-1,\ 1 \leqq i \leqq r_\nu\}$ be linearly independent. Then, $\langle A, B\rangle$ is C-equivalent with a system $\langle A', B'\rangle$, where A', B' have the form of lemma 1 with α's and γ's time-dependent.*

In essence, this lemma is proved in [4], but we shall give an alternate proof, which is modeled after the proof of lemma 2 of this paper from [6].

Proof. Let $P\langle A, B\rangle = \{p_i | 1 \leqq i \leqq m\}$, $k_i = \sum_{\nu=1}^{i} p_\nu$ and let $\tilde{C}$ be the submatrix of C, consisting of its k_i-th columns, G the submatrix of B', consisting of its k_i-th rows. Then, C, G have to satisfy

$$AC = CA' + \dot{C}, \quad B = \tilde{C}G, \tag{12}$$

G being triangular, so is G^{-1}. We define $G^{-1} = (\gamma_{\nu i})$ as follows: $\gamma_{\nu i}$ for $1 \leqq \nu < r_{p_i}$ are the unique nunbers such that

$$\pi_{p_i}(\mathfrak{A}^{p_i} b_i + \sum_{\nu=1}^{r_{p_i}} \gamma_{\nu i} \mathfrak{A}^{p_i} b_\nu) = 0,$$

$\gamma_{ii} = 1$ and all remaining γ's are equal to zero.

184 From lemma 3 it follows that γ_{vi} are $\mathscr{C}^\infty$. Since $\tilde{C} = BG^{-1}$, we have $\pi_{p_i}(\mathfrak{A}^{p_i}c_{k_i}) = \pi_{p_i}(\mathfrak{A}^{p_i}b_i - \sum_{v=1}^{r_{pi}} \gamma_{vi}\mathfrak{A}^{p_i}b_v) = 0$, or $\mathfrak{A}^{p_i}c_{k_i} \in L_{p_i-1}$. Moreover,

$$\{\mathfrak{A}^v c_{k_i} | 0 \leqq v \leqq j\,, \quad 1 \leqq i \leqq r_v\} \tag{13}$$

are bases of L_j.

Decomposing the first equality of (12) into columns, we obtain

$$c_{k_{i-1}+j} = \mathfrak{A}c_{k_{i-1}+j+1} - \sum_{v=1}^{r_0} \alpha_{vk_{i-1}+j+1}c_{k_v}\,, \quad j = 1, \ldots, p_i - 1,\ i = 1, \ldots, m\,,$$

$$0 = \mathfrak{A}c_{k_{i-1}+1} - \sum_{v=1}^{r_0} \alpha_{vk_{i-1}+1}c_{k_v}\,, \quad i = 1, \ldots, m\,.$$

Consequently

$$c_{k_{i-1}+j} = \mathfrak{A}^{p_i-j}c_{k_i} - \sum_{\mu=1}^{p_i-j} \mathfrak{A}^{p_i-j-\mu} \sum_{v=1}^{r_0} \alpha_{vk_i-\mu+1}c_{k_v}\,, \tag{14}$$

$$0 = \mathfrak{A}^{p_i}c_{k_i} - \sum_{\mu=1}^{p_i} \mathfrak{A}^{p_i-\mu} \sum_{v=1}^{r_0} \alpha_{vk_i-\mu+1}c_{k_v}\,. \tag{15}$$

Since for any $\mathscr{C}^\infty$ $\alpha(t)$ scalar and $b(t)$ n-vector functions

$$\mathfrak{A}^j(\alpha b) = \sum_{\mu=0}^{j} \binom{j}{\mu} (-1)^{j-\mu} \frac{\mathrm{d}^{j-\mu}\alpha}{\mathrm{d}t^{j-\mu}}\, \mathfrak{A}^\mu b \tag{16}$$

is valid, we can re write (15) as

$$\mathfrak{A}^{p_i}c_{k_i} = \sum_{j=0}^{p_i-1} \sum_{v=1}^{r_0} \sum_{\mu=0}^{j} \binom{j}{\mu} (-1)^{j-\mu}\, \alpha^{(j-\mu)}_{vk_{i-1}+j+1}\, \mathfrak{A}^\mu c_{k_v}$$

$$= \sum_{\mu=0}^{p_i-1} \sum_{\mu=1}^{r_0} \mathfrak{A}^\mu c_{k_v} \sum_{j=\mu}^{p_i-1} \binom{j}{\mu} (-1)^{j-\mu}\, \alpha^{(j-\mu)}_{vk_{i-1}+j+1}\,.$$

If we define $\varphi_{v\mu} \equiv 0$ for $\mu \geqq p_v$, then by (13) $\varphi_{v\mu}$, $1 \leqq v \leqq r_0$, $0 \leqq \mu \leqq p_i - 1$ are uniquely determined by the equation

$$\mathfrak{A}^{p_i}c_{k_i} = \sum_{\mu=0}^{p_i-1} \sum_{v=1}^{r_0} \mathfrak{A}^\mu c_{k_v}\varphi_{v\mu}$$

and they are $\mathscr{C}^\infty$ on J.

$\varphi_{v\mu}$ being known, $\alpha_{v\zeta}$ can be determined by solving r_0 triangular systems with 1's in the diagonal

$$\sum_{j=\mu}^{p_i-1} \binom{j}{\mu} (-1)^{j-\mu}\, \alpha^{(j-\mu)}_{vk_{i-1}+j+1} = \varphi_{v\mu}\,.$$

Obviously, the α's obtained from these equations are also $\mathscr{C}^\infty$ on J. The α's being known, we can determine C by (14) and it can be readily verified that (12) is satisfied.

Since by (14) and (16), $c_{k_{i-1}+j} = \mathfrak{A}^{p_i-1-j}c_{k_i} + f$, where $f \in L_{p_i-j}$, it follows from (13) that C is nonsingular. 185

Remark 2. It can be easily checked that if we want the α's and γ's to be merely continuous, it is sufficient to assume that $(B, \mathfrak{A}B, \ldots, \mathfrak{A}^{n-1}B)$ has $\max p_i - 1$ derivatives. Moreover, if these derivatives are bounded and $|\det S(t)|$, where $S(t)$ is the matrix, consisting of the columns $\{\mathfrak{A}^v b_i | 0 \leqq v \leqq n-1,\ 1 \leqq i \leqq r\}$ is bounded below by a positive constant, C, C^{-1} and A', B' are also obtained bounded.

Proof of theorem 3. If: By lemma 7 and 8, $\langle A, B\rangle$ can be brought to $\langle A', B'\rangle$ with A', B' having the special form of lemma 2 with α's and γ's time dependent. In the same way as in theorem 2, we construct Q (which will be, of course, time dependent) in such a way that $\langle A', B'\rangle$ will be Q-equivalent to the decoupled system of r_0 integrators (6), which is autonomous.

Only if: Lemma 5 proves that the r-numbers are invariants of a D-transformation. The theorem will be proved if we show that they are also invariant of the C- and Q-transformation.

For the C-transformation if follows from $\mathfrak{A}'f' = C^{-1}(AC - \dot{C})C^{-1}f - \dot{C}^{-1}f + C^{-1}\dot{f} = C^{-1}\mathfrak{A}f$ for every $\mathscr{C}^\infty$ vector function f, where $\langle A', B'\rangle$ is the C-transform of $\langle A, B\rangle$ and $\mathfrak{A}' = (A' - \mathrm{d}/\mathrm{d}t)$.

For the F-transformation we note that if $\langle A', B\rangle$ is a Q-transform of $\langle A, B\rangle$, then $\mathfrak{A}'f = (A + BQ)f - \dot{f} = \mathfrak{A}f + BQf$, from which it follows $L_j < A', B > \subset \subset L_j\langle A, B\rangle$. From the symmetry of Q-equivalence it follows $L_j\langle A, B\rangle \subset L_j\langle A', B\rangle$ and, thus, $L_j\langle A, B\rangle = L_j\langle A', B\rangle$, q. e. d.

Those systems, which satisfy the assumptions of theorem 3 we shall call autonomous-equivalent, or A-systems.

We have also proved

Theorem 4. *Two A-systems $\langle A, B\rangle$ and $\langle A', B'\rangle$ are F-equivalent if and only if $R\langle A, B\rangle = R\langle A', B'\rangle$. They are both equivalent to the canonical system* (6).

Corollary 4. *For autonomous systems, we obtain the same classification, whether we allow the transformation matrices to be time-dependent or not.*

Corollary 5. *To any n-th order polynomial $P(\lambda)$ and any A-system $\langle A, B\rangle$ there is an autonomous system $\langle A', B'\rangle$, F-equivalent with $\langle A, B\rangle$ such that $P(\lambda)$ is the characteristic polynomial of A'* (cf. [4]).

Remark 3. Corollary 5 is of use for the stabilization problem only if $J = [t_0, \infty)$ and the transformation matrices and the inverses of the C- and D-matrices are bounded. In virtue of remarks 1 and 2, this will be true if the assumptions of remark 1 are satisfied on some interval $J = [t_0, \infty)$. Let us also note that for the stabilization problem solely, it is sufficient to assume that the matrix $(B, \mathfrak{A}B, \ldots, \mathfrak{A}^{p-1}B)$, $p =$

186 $= \max_{1 \leqq i \leqq m} p_i$ has $n - 1$ bounded continuous dérivatives. Our results, therefore, improve the result of [4], where it is in essence assumed that the system satisfies the assumptions of Remark 2. We assume that A, B are $\mathscr{C}^\infty$ merely for the classification theory. Namely, if $\langle A, B\rangle$ is not assumed to be $\mathscr{C}^\infty$, then the system $\langle A', B'\rangle$ of lemma 8 is obtained with much fewer continuous derivatives, which is inconvenient.

Corollary 6. *Let $\langle A, B\rangle$ be an A-system and let $f(t)$ bea $\mathscr{C}^\infty$ vector function such that for all $t \in J$, $f(t) \in L_0\langle A, B\rangle(t)$, Then, there is a system $\langle A', B'\rangle$, F-equivalent with $\langle A, B\rangle$, such that $\langle A', f\rangle$ is an A-system.*

The idea of proof is the same as that of the proof of Corollary 2; we omit the details.

4. REMARKS ON DISCRETE AND PERIODIC SYSTEMS

For discrete systems, autonomous as well as time-dependent, a similar classification theory can be developed. We are not going to formulate the results, which can be drawn from those of 2–3 by simple analogies.

Time-varying periodic systems with continuous time can be regarded as a particular case of general time-varying systems. It is natural to require from the transformation matrices to be periodic in this case but this requirement does not introduce new difficulties.

Sometimes, however, one does not need the information about the behaviour of the system for all t, but only about its behaviour in discrete moments, the period of the system apart. For instance the stability properties of the system are completely determined by the discretized system.

To make this point more precise, assume that the matrices of the system $\langle A, B\rangle$ are T-periodic and of class $\mathscr{C}^1$. The solutions of the system $\dot{x} = Ax + Bu$ can be expressed as

$$x(t) = Y(t)\,x(0) + \int_0^t Y(t)\,Y(-s)\,B(s)\,u(s)\,\mathrm{d}s$$

where $Y(t)$ is the fundamental matrix of $\dot{y} = Ay$ with initial condition $Y(0) = E$. Then, the corresponding discrete system is

$$\xi_{k+1} = Y(T)\,\xi_k + F(\eta_k)$$

where $\xi_k = x(kT)$, η_k is the piece of control function $u(s)$, $kT \leqq s \leqq (k+1)T$ and F is the linear operator, mapping η_k into

$$\int_0^T Y(T)\,Y(-s)\,B(0)\,u(kT + s)\,\mathrm{d}s\,.$$

It can be immediately seen that if we try to use the concept of D-transformation

to this system, difficulties arise. Namely, a general linear transformation in the η-space would involve the future values of the control, which is physically unthinkable. 187

However, there is a result of the type of corrollary 1, which may be worthwhile to mention in this context:

Theorem 5. *A T-periodic system* $\langle A, B\rangle$ *with* $\langle A, B\rangle$ *bein* $\mathscr{C}^1$ *is controllable if and only if to every n-th degree polynomial* $P(\lambda)$ *with positive absolute term there is a T-periodic* $m \times n$ *matrix* Q *(which can be choosen piecewise constant or arbitrarily smooth) such that the characteristic multipliers of the system* $\dot{y} = (A + BQ)\, y$ *are exactly the roots of* $P(\lambda)$. *In particular, periodic controllable systems can always be stabilized by an appropriate periodic feedback.*

This theorem is in a sense stronger than corollary 4, since controllability in it is assumed only in its weakest geometrical sense, i. e. that we can join any two points in R^n by a trajectory of the system in sufficiently long time, with the aid of an appropriate control.

For the proof of theorem 5, see [7].

(Received April 2, 1969.)

REFERENCES

[1] V. M. Popov: Hyperstability and optimality of automatic systems with several control functions. Rev. Roumaine Sci. Tech., Electrotechn. et Energ. *9* (1964), 629—690.
[2] V. M. Popov: Hyperstabilitatea sistemelor automate. Editura Academiei Rep. Soc. Romania, Bucharest 1966.
[3] E. G. Gilbert: The decoupling of multivariable systems by state feedback. SIAM Journal on Control *7* (1969), 50—63.
[4] W. A. Wolowich: On the stabilization of controllable systems. (Pre-print.)
[5] R. E. Kalman: Algebraic aspects of the theory of dynamical systems. Differential equations and dynamical systems, ed. by J. K. Hale and J. P. LaSalle, Academic Press 1967, 133—146.
[6] П. Бруновскы: О стабилизации линейных систем при определенном классе постоянно действующих возмущений. Дифференциальные уравнения 2 (1966), 769—777.
[7] P. Brunovský: Controllability and linear closed-loop controls in linear periodic systems. Journal of Differential equations *6* (1969), 296—313.
[8] C. E. Langenhop: On the stabilization of linear systems. Proc. Am. Math. Soc. *15* (1964), 735—742.
[9] W. M. Wonham: On pole assignment in multi-input controllable linear systems. IEEE Transactions on automatic control *AC-12* (1967), 660—665.
[10] G. H. Hardy, E. M. Wright: An Introduction to the theory of numbers. Clarendon 1938.
[11] V. Doležal: The existence of a continuous basis of a certain linear subspace of E_r which depends on a parameter. Čas. pěst. mat. *89* (1964), 466—469.

188 VÝŤAH

Klasifikácia lineárnych riaditeľných sústav

PAVOL BRUNOVSKÝ

Vyšetruje sa relácia F-ekvivalencie a na nej založená klasifikácia lineárnych riaditeľných sústav. Dve lineárne sústavy riadenia $\dot{x} = Ax + Bu$ a $\dot{y} = Ay + Bv$ (A, B, A', B' konštantné), spĺňujúce predpoklad riaditeľnosti (2) sa nazývajú F-ekvivalentné, ak existujú matice Q, C, D (C, D regulárne) také, že je splnené (3). Odvozuje sa nutná a postačujúca podmienka ekvivalencie sústav, spočívajúce v rovnosti konečného počtu čísel, zviazaných s maticami A, B, resp. A', B'. Ukazuje sa, že v každej triede ekvivalencie existuje kanonická sústava, pozostávajúca z nezávislých integrátorov, ktorých počet a rády úplne charakterizujú triedu ekvivalencie.

Pojem F-ekvivalencie sa zovšeobecňuje pre sústavy závislé od času a to tak, že sa povoľujú matice Q, C, D závislé od času. Vyšetruje sa trieda sústav, F-ekvivalentných s časovo nezávislými sústavami – sú to práve tie, ktoré splňujú podmienku (9).

Nakoniec sa stručne diskutuje prípad sústav periodických a diskrétnych v čase.

RNDr. Pavol Brunovský CSc., Ústav technickej kybernetiky SAV, Dúbravská cesta, Bratislava 9.

P. Brunovský

Local controllability of odd systems

Banach Center Publications, Vol. 1 (1976), 39–45.

BANACH CENTER PUBLICATIONS
VOLUME 1

LOCAL CONTROLLABILITY OF ODD SYSTEMS

PAVOL BRUNOVSKÝ

Mathematical Institute, Slovak Academy of Sciences, Bratislava, Czechoslovakia

1. Introduction

Consider a control system

$$\dot{x} = f(x, u), \tag{1}$$

where $x \in \boldsymbol{R}^n$ and $u \in U$ (we do not specify the set U at this point) and a set of admissible controls $\mathcal{U}$ which is a subset of the set of mappings u of intervals $[0, T(u)]$, $T(u) \geqslant 0$ into U, having the property that for any $u \in \mathcal{U}$ the solution $\varphi(t, x_0, u)$ of the differential equation

$$\dot{x} = f(x, u(t))$$

with $\varphi(0, x_0, u) = x_0$ is uniquely defined on $[0, T(u)]$. We denote by $\mathcal{R}(x_0)$ the reachable set of (1) from x_0, i.e., $\mathcal{R}(x_0) = \{\varphi(T(u), x_0, u) | \ u \in \mathcal{U}\}$ and call (1) *locally controllable at* x_0 if $x_0 \in \operatorname{int} \mathcal{R}(x_0)$.

The well-known theorem of Kalman gives necessary and sufficient conditions of local controllability at 0 for the class of linear systems $(f(x, u) = Ax + Bu)$ with U being a subset of R^m containing 0 in its interior (the bang-bang controllability theorem makes the choice of $\mathcal{U}$ irrevelant in this case). For nonlinear systems, sufficient conditions for local controllability at a rest point of the uncontrolled system are given by the theorem of Lee and Markus ([4]). However, since the theorem of Lee and Markus uses only the linearization of the system (in both x and u), it is not difficult to see that its sufficient condition is far from being necessary.

A more recent approach, going back to Hermann ([2], cf. also [5], [6]) relates the problem of controllability to the study of orbits of families of vector fields. Given a family of vector fields $\mathcal{X} = \{X^i | \ i \in I\}$ on R^n, the orbit of x_0 is defined as $\Omega(x_0) = \{\varphi^{i_p}_{t_p} \circ \ldots \circ \varphi^{i_1}_{t_1}(x_0) | \ p \geqslant 0, i_j \in I, t_j \in R, j = 1, \ldots, p\}$, where by φ^i we denote the flow of X^i. It is immediately seen that if $\mathcal{U}$ is taken as the set of piecewise constant controls and we associate with (1) the family of vectors fields $\mathcal{X} = \{X^u | \ u \in U\}$ defined by $X^u(x) = f(x, u)$, then $\mathcal{R}(x_0) = \Omega^+(x_0)$, were $\Omega^+(x_0)$ is the positive semiorbit of x_0, defined by $\Omega^+(x_0) = \{\varphi^{u_p}_{t_p} \circ \ldots \circ \varphi^{u_1}_{t_1}(x_0) | \ p \geqslant 0, u_i \in U, t_i \geqslant 0, i = 1, \ldots, p\}$.

It is not true in general that $\Omega^+(x_0) = \Omega(x_0)$. The only simple (but rather restrictive) condition guaranteeing this is the symmetry condition: for every x and every $i \in I$ there exists a $j \in I$ such that $X^j = -X^i$ in some neighbourhood of x (cf. [5]).

For families of analytic vector fields, the classical theorem of Chow gives a certain rank condition (cf. Theorem 1 below), which is necessary and sufficient for $x_0 \in \operatorname{int} \Omega(x_0)$ and which, if applied to linear systems, is equivalent to the rank condition of Kalman (cf. [2], [5]). However, Chow's theorem does not yield a generalization of Kalman's one since the family of vector fields, associated with a linear system is not symmetric in general.

The aim of this paper is to prove the equivalence of Chow's rank condition to local controllability for systems exhibiting a different kind of symmetry which is satisfied for linear systems—a theorem which does contain Kalman's controllability theorem as its special case.

In § 2, we formulate the main theorem in the language of families of vector fields and three lemmas from which the proof of the theorem easily follows. The applications of the main theorem to local controllability of control systems are given in § 3 and § 4 contains the proof of Lemma 3.

2. Main theorem

We shall call a family of vector fields $\mathscr{X} = \{X^i | \, i \in I\}$ *odd* if for every $i \in I$ there exists a $j \in I$ such that $X^i(-x) = -X^j(x)$ for all x. An odd family of vector fields can always be indexed in such a way that I contains symbols $+i$ and $-i$ in such a way that $X^{-i}(x) = -X^i(-x)$ (sometimes X^i and X^{-i} may coincide). When dealing with an odd family of vector fields we shall always assume that it is indexed in this way.

Further, we shall always assume that all the vector fields under consideration are complete, i.e., that the domain of existence of their integral curves is R. This assumption, just as the assumption that the vector fields are defined and satisfy the oddness assumption over all $\boldsymbol{R}^n$ (instead of a neighbourhood of 0) is not essential and is made only for the sake of simplicity.

With a family $\mathscr{X}$ of C^∞ vector fields we associate the family $[\mathscr{X}]$, which is the smallest family of vector fields containing $\mathscr{X}$ and closed under the formation of Lie brackets (cf. [2], [5], [6]). We write $\mathscr{X}(x) = \{X(x) | \, X \in \mathscr{X}\}$.

THEOREM 1. *Let $\mathscr{X}$ be an odd family of analytic vector fields on R^n. Then $0 \in \operatorname{int} \Omega^+(0)$ if and only Chow's rank condition is satisfied at 0, i.e.,* $\dim \operatorname{span} [\mathscr{X}](0) = n$.

For the proof we need the following three lemmas.

LEMMA 1. *Let $\mathscr{X} = \{X^i | \, i \in I\}$ be a family of C^∞ vector fields on $\boldsymbol{R}^n$ satisfying Chow's rank condition at x_0. Then, for every $\delta > 0$, there exist $i_1, \ldots, i_n \in I$, $s_1, \ldots, s_n \in [0, \delta)$ such that the map $(t_1, \ldots, t_n) \mapsto \varphi^{i_n}_{t_n} \circ \ldots \circ \varphi^{i_1}_{t_1}(x_0)$ is a local diffeomorphism at $(s_1, \ldots, s_n)$.*

For the proof, cf. [3].

LEMMA 2. *Let* $\mathscr{X} = \{X^i \mid i \in I\}$ *be a family of* C^∞ *vector fields,* $y \in \operatorname{int} \Omega^+(x)$, $z \in \Omega^+(y)$. *Then,* $z \in \operatorname{int} \Omega^+(x)$.

Proof. From $z = \varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_1}^{i_1}(y)$ it follows that

$$z \in \varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_1}^{i_1}(\operatorname{int} \Omega^+(x)) \subset \operatorname{int}\left(\varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_1}^{i_1}(\Omega^+(x))\right) \subset \operatorname{int} \Omega^+(x),$$

since $\varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_1}^{i_1}$ is a local diffeomorphism.

To make the formulation of Lemma 3 easier, we define for a given family of analytic vector fields, a *stream* on $V \subset \boldsymbol{R}^n$ open as an analytic map $\chi\colon (-\delta, \delta) \times \times V \to \boldsymbol{R}^n$, $\delta > 0$ (write $\chi_t(x) = \chi(t, x)$) such that

1. for every $t \in (-\delta, \delta)$, χ_t is a diffeomorphism $V \to \chi_t(V)$,
2. for all $x \in V$, $\chi_0(x) = x$,
3. for every $t \in [0, \delta)$ and all $x \in V$, $\chi_t(x) \in \Omega^+(x)$.

Note that if χ, ψ are streams on V and τ_1, τ_2 are analytic functions on a neighbourhood of 0 such that

$$\tau_1(0) = \tau_2(0) = 0 \quad \text{and} \quad \tau_1(t) > 0,\ \tau_2(t) > 0 \text{ for } t > 0, \tag{2}$$

then $t \mapsto \chi_{\tau_1(t)} \circ \psi_{\tau_2(t)}$ is also a stream on V. If for two streams χ, ψ there exists a stream η and analytic functions τ_1, τ_2 satisfying (2) such that $\psi_t = \eta_{\tau_1(t)} \circ \chi_{\tau_2(t)}$ for $|t|$ sufficiently small, we shall write $\chi \prec \psi$. The relation $\prec$ is obviously transitive.

LEMMA 3. *Let* $\mathscr{X}$ *be an odd family of analytic vector fields. Then for every stream* χ *there exists a stream* $\vartheta \succ \chi$ *such that* $\vartheta_t(0) = 0$ *for* $t \geqslant 0$ *sufficiently small.*

Proof of Theorem 1. *Sufficiency.* Write $\chi_t(x) = \varphi_{ts_n}^{i_n} \circ \ldots \circ \varphi_{ts_1}^{i_1}(x)$, where $i_1, \ldots, i_n, s_1, \ldots, s_n$ are chosen as in Lemma 1. Obviously, χ is a stream on some neighbourhood of 0. The Jacobian of χ_t at 0 is an analytic function of t which does not vanish for $t = 1$. Therefore, it must be non-zero for $t > 0$ sufficiently small. Consequently, $\chi_t(0) \in \operatorname{int} \Omega^+(0)$ for $t > 0$ sufficiently small. Let ϑ be as in Lemma 3. Then there exist analytic functions τ_1, τ_2 satisfying (2) and a stream η such that $\eta_{\tau_2(t)} \circ \chi_{\tau_1(t)}(0) = 0$ (which implies $0 \in \Omega^+(\chi_t(0))$) for $t > 0$ sufficiently small. By Lemma 2, $0 \in \operatorname{int} \Omega^+(0)$.

The *necessity* of Chow's condition follows from the fact that $\Omega^+(0) \subset \Omega(0)$ and that, if Chow's condition is not satisfied, $\Omega(0)$ is a submanifold of $\boldsymbol{R}^n$ of dimension $< n$ (cf. [5], [6]).

Let us note that although Theorem 1 is formulated in $\boldsymbol{R}^n$, its nature is local. Thus, we can replace $\boldsymbol{R}^n$ by an n-dimensional analytic manifold, provided the oddness assumption is satisfied in some local chart at 0. This is the situation if e.g. Chow's condition is not satisfied and we consider the restriction of $\mathscr{X}$ to the orbit $\Omega(0)$, which is an analytic submanifold of R^n of dimension $< n$ (note that $\Omega(0)$ is symmetric with respect to 0 if $\mathscr{X}$ is odd!); cf. [5], [6]. Thus we have

THEOREM 2. *Let* $\mathscr{X}$ *be an odd family of analytic vector fields. Then* $0 \in \operatorname{int} \Omega^+(0)$ *in the topology of* $\Omega(0)$.

3. Application to control systems

Consider a control system

$$\dot{x} = f_0(x) + \sum_{i=1}^{p} u_i f_i(x), \quad u_i \in U_i = [-1, +1], \tag{3}$$

$U = U_1 \times U_2 \times \ldots \times U_p$. We associate with (3) the family of vector fields $\mathscr{X} = \{f_0 \pm f_i | i = 1, \ldots, p\}$. If we take as $\mathscr{U}$ the set of piecewise constant bang-bang controls (i.e., the set of piecewise constant controls with values ± 1), then obviously $\mathscr{R}(x) = \Omega^+(x)$. Let us also note that since $\mathscr{X}(x)$ and $\{f_i | i = 0, \ldots, p\}(x)$ span the same linear subspace, so do $[\mathscr{X}](x)$ and $[\{f_i | i = 0, \ldots, p\}](x)$. Thus we obtain the following corollary of Theorem 1:

THEOREM 3. *Let f_i, $i = 0, \ldots, p$ be analytic, let f_0 be odd, and let f_i, $i = 1, \ldots, p$, be odd or even. Then* (3) *is locally controllable at* 0 *if and only if* $\operatorname{rank}[\{f_i | i = 0, \ldots, p\}](0)$ *is* n.

We omit the obvious reformulation of Theorem 2 in the language of control systems.

Let us note that the conditions of Theorem 3 are satisfied if f_0 is linear and f_i, $i = 1, \ldots, p$, are constant and so Kalman's controllability theorem is obtained as a special case of Theorem 3.

The perturbation theory of [1] allows us to extend the controllability result of Theorem 3 to "allmost odd" control systems:

THEOREM 4. *Given a system* (3) *satisfying the assumptions of Theorem* 3 *such that* $\dim \operatorname{span}[\{f_i | i = 0, \ldots, p\}] = n$, *there exist* $\varepsilon > 0$ and $\eta > 0$ *such that for any function* $g(x, u)$ *which is Lipschitz continuous in* x *and continuous in* u *and satisfies* $|g(x, u)| < \varepsilon$ *for* $|x| < \eta$ *the system*

$$\dot{x} = f_0(x) + \sum_{i=1}^{p} u_i f_i(x) + g(x, u)$$

is locally controllable at 0.

The proof follows from [1], Proposition III-6. One has merely to note that the homogenity assumption is not essential in this proposition.

4. Proof of Lemma 3

For the sake of brevity we make the following convention: By a stream we shall always understand a stream on some neighbourhood of the origin. In statements concerning t we shall drop "for $|t|$ sufficiently small".

Let us note that if $\mathscr{X}$ is odd, for any stream χ the symmetric map χ^- defined by $\chi_t^-(x) = -\chi_t(-x)$ is also a stream. For the proof it suffices to note that $\chi_t(x) = \varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_2}^{i_1}(x)$ implies

$$\chi_t^-(x) = -\chi_t(-x) = -\varphi_{t_p}^{i_p} \circ \ldots \circ \varphi_{t_1}^{i_1}(-x) = \varphi_{t_p}^{-i_p} \circ \ldots \circ \varphi_{t_1}^{-i_1}(x) \in \Omega^+(x).$$

In the sequel we shall always denote pairs of symmetric streams by the same letter with superscripts $+$, $-$, sometimes dropping $+$. When dealing with them simultaneously we shall use the letter δ to indicate the signs; if multiplied, $+$, $-$ will be understood to behave like $+1$, -1.

In order to prove Lemma 3 we prove the following induction statement:

Let χ_i, $i = 1, \ldots, k$, ψ_k *be streams such that*

$$(4_k) \qquad \chi_{i,t}(0) = a_i t^{p_i} + o(t^{p_i+1}), \quad \psi_{k,t}(0) = b_k t^{q_k} + o(t^{q_k+1}),$$

where a_i, $i = 1, \ldots, k$, *are linearly independent and* b_k *does not belong to any subspace spanned by* $k-1$ *of the vectors* a_i, $i = 1, \ldots, k$.

Then either there exists a stream $\psi_{k+1} \succ \psi_k$ *such that* $\psi_{k+1,t}(0) = 0$ *or there exist streams* χ_{k+1}, ψ_{k+1} *such that* $\psi_{k+1} \succ \psi_k$ *and* χ_i, $i = 1, \ldots, k+1$, *and* ψ_{k+1} *satisfy* (4_{k+1}).

The assertion of the lemma results from this induction statement as follows: If $\chi_t(0) \equiv 0$, we write $\vartheta = \chi$. Otherwise, (4_1) is satisfied for $\chi_1 = \psi_1 = \chi$. Using the induction statement we construct à sequence of streams $\psi_1 \prec \psi_2 \prec \ldots$ (and the auxiliary streams $\chi_1, \chi_2, \ldots$) until we reach k_0 such that $\psi_{k_0,t}(0) = 0$ ând we write $\vartheta = \psi_{k_0}$. Since (4_{n+1}) is impossible, $k_0 \leqslant n+1$.

To prove the induction statement we write

$$\xi_{i,t} = \chi_{i,s_i(t)}, \quad \text{where} \quad s_i(t) = t^{p_1 \ldots p_{i-1} p_{i+1} \ldots p_k q_k},$$

$$\eta_{k,t} = \psi_{k,r_k(t)}, \quad \text{where} \quad r_k(t) = t^{p_1 \ldots p_k},$$

ξ_i, $i = 1, \ldots, k$, and η_k are streams, $\xi_i \succ \chi_i$, $\eta_k \succ \psi_k$ and

$$\xi_{i,t}(0) = a_i t^Q + O(t^{Q+1}),$$

$$\eta_{k,t}(0) = b_k t^Q + O(t^{Q+1}),$$

where $Q = p_1 \ldots p_k q_k$. Further we have

$$(5) \qquad \begin{aligned} \xi_{i,t}(x) &= x + \sum_{j=1}^{Q} \alpha_{ij}(x) t^j + O(t^{Q+1}), \\ \eta_{k,t}(x) &= x + \sum_{j=1}^{Q} \beta_j(x) t^j + O(t^{Q+1}), \end{aligned}$$

where $\alpha_{ij}(x) = O(|x|)$, $\beta_j(x) = O(|x|)$ for $j = 0, \ldots, Q-1$ and $\alpha_{iQ}(x) = a_i + O(|x|)$, $\beta_Q(x) = b_k + O(|x|)$.

Assume that there exists no stream $\psi_{k+1} \succ \psi_k$ such that $\psi_{k+1,t}(0) = 0$. Write $T_k = \operatorname{span}\{a_i |\ i = 1, \ldots, k\}$ and choose such a complement S_k to T_k that if π_k denotes the projection onto T_k along S_k, then $\pi_k(b_k)$ does not lie in any subspace spanned by $k-1$ of the vectors a_i, $i = 1, \ldots, k$. This is possible owing to the assumption that b_k itself does not belong to any such subspace. We show that there exist functions $\tau_1, \ldots, \tau_k$ and signs $\delta_1, \ldots, \delta_k$ such that $\tau_i(0) = 0$, $\tau_i(0) > 0$ for $t > 0$, $i = 1, \ldots, k$, and

$$(6) \qquad \pi_k \circ \xi^{\delta_1}_{1,\tau_1(t)} \circ \ldots \circ \xi^{\delta_k}_{k,\tau_k(t)} \circ \eta_{k,t}(0) = 0.$$

Denote $F^{\delta_1,\dots,\delta_k}\colon R^{n+1} \to T_k$ by

$$F^{\delta_1,\dots,\delta_k}(\tau_1, \dots, \tau_k, t) = \pi_k \circ \xi^{\delta_1}_{1,\tau_1} \circ \dots \circ \xi^{\delta_k}_{k,\tau_k} \circ \eta_{k,t}(0).$$

By (5) we have

$$F^{\delta_1,\dots,\delta_k}(\tau_1, \dots, \tau_k, t) = \sum_{i=1}^{k} \delta_i a_i \tau_i^Q + \pi_k(b_k) t^Q + \omega(\tau_1, \dots, \tau_k, t),$$

where ω is analytic and satisfies

$$\omega(\tau_1, \dots, \tau_k, t) = o(|\tau_1|^Q + \dots + |\tau_k|^Q + |t|^Q). \tag{7}$$

Write

$$G^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = F^{\delta_1,\dots,\delta_k}(\sigma_1 t, \dots, \sigma_k t, t).$$

Then we have

$$G^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = t^Q \Big[\sum_{i=1}^{k} \delta_i a_i \sigma_i^Q + \pi_k(b_k)\Big] + \omega(\sigma_1 t, \dots, \sigma_k t, t).$$

By (7) we have $\partial^{j_1+\dots+j_{k+1}}\omega(\sigma_1 t, \dots, \sigma_k t, t)/\partial\sigma_1^{j_1} \dots \partial\sigma_k^{j_k}\partial t^{j_{k+1}}(0) = 0$ as soon as $j_{k+1} \leqslant Q$. Thus, by the Weierstrass preparation theorem, $\omega(\sigma_1 t, \dots, \sigma_k t, t) = t^{Q+1}\tilde{\omega}(\sigma_1, \dots, \sigma_k, t)$, where $\tilde{\omega}$ is analytic in $\sigma_1, \dots, \sigma_k, t$. Therefore,

$$G^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = t^Q \Big[\sum_{i=1}^{k} \delta_i a_i \sigma_i^Q + \pi_k(b_k) + t\tilde{\omega}(\sigma_1, \dots, \sigma_k, t)\Big]$$

and

$$G^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = 0 \quad \text{if} \quad H^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = 0,$$

where

$$H^{\delta_1,\dots,\delta_k}(\sigma_1, \dots, \sigma_k, t) = \sum_{i=1}^{k} \delta_i a_i \sigma_i^Q + \pi_k(b_k) + t\tilde{\omega}(\sigma_1, \dots, \sigma_k, t).$$

Since a_i, $i = 1, \dots, k$, form a basis of T_k and $\pi_k(b_k)$ does not belong to any subspace spanned by $k-1$ of the vectors a_i, there exists a unique k-tuple of reals γ_i, all of them $\neq 0$, such that $-\pi(b_k) = \sum_{i=1}^{k} a_i\gamma_i$. We write $\delta_i = \operatorname{sign}\gamma_i$ and choose $\sigma_i^* = (\delta_i\gamma_i)^{1/Q}$. Since $H^{\delta_1,\dots,\delta_k}(\sigma_1^*, \dots, \sigma_k^*, 0) = \sum_{i=1}^{k} a_i\gamma_i + \pi_k(b_k) = 0$ and $\partial H^{\delta_1,\dots,\delta_k}/\partial\sigma\,(\sigma_1^*, \dots, \sigma_k^*, 0) = Q(\delta_1\,\sigma_1^{*Q-1}a_1, \dots, \delta_k\sigma_k^{*Q-1}a_k)$ is nonsingular (here $\sigma = (\sigma_1, \dots, \sigma_k)$), there exists a unique k-tuple of analytic functions $\sigma_i(t)$ such that $\sigma_i(0) = \sigma_i^*$ and $H^{\delta_1,\dots,\delta_k}(\sigma_1(t), \dots, \sigma_k(t), t) = 0$. Moreover, $\sigma_i(t) \geqslant 0$. If we write $\tau_i(t) = t\sigma_i(t)$, $i = 1, \dots, k$, then τ_i will be analytic and will satisfy $\tau_i(0) = 0$, $\tau_i(t) > 0$ for $t > 0$ and $F^{\delta_1,\dots,\delta_k}(\tau_1(t), \dots, \tau_k(t), t) = 0$.

Write

$$\chi_{k+1,t}(x) = \xi^{\delta_1}_{1,\tau_1(t)} \circ \dots \circ \xi^{\delta_k}_{k,\tau_k(t)} \circ \eta_{k,t}.$$

Obviously $\chi_{k+1} \succ \psi_k$; thus $\chi_{k+1,t}(0) \neq 0$ by assumption. By (6) and the definition of χ_{k+1}, $\pi_k \circ \chi_{k+1,t}(0) = 0$, which implies that the first non-zero coefficient in the

expansion of $\chi_{k+1,t}(0)$ must be linearly independent of the vectors a_i, $i = 1, \ldots, k$. Therefore, $\chi_1, \ldots, \chi_{k+1}$ satisfy (4_{k+1}).

To obtain ψ_{k+1} we choose another complement S'_k to T_k, the intersection of which with the span $\{a_i | \ i = 1, \ldots, k+1\}$ does not lie in any subspace spanned by k of the vectors a_i, $i = 1, \ldots, k+1$. We construct ψ_{k+1} by the same construction as χ_{k+1} with S_k replaced by S'_k and π'_k replaced by π'_k, the projection onto T_k along S'_k. Since $\psi_{k+1,t}(0) \neq 0$, owing to the choice of S'_k, χ_i, $i = 1, \ldots, k+1$, and ψ_{k+1} will satisfy (4_{k+1}).

References

[1] P. Brunovský, C. Lobry, *Contrôlabilité bang bang, contrôlabilité differentiable et perturbations des systèmes non linéaires*, to appear in Annali Mat. Pura Appl.

[2] R. Hermann, *On the accessibility problem in control theory*, in Internat. Symp. on Nonlin. Dif. Eq. and Nonlin. Mech., J. P. LaSalle and S. Lefschetz editors, Academic Press, New York 1963, pp. 325–332.

[3] A. Krener, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear systems*, SIAM J. Control 12 (1974), No. 1.

[4] E. B. Lee, L. Markus, *Foundations of optimal control theory*, Wiley, New York 1967.

[5] C. Lobry, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control 8 (1970), pp. 573–605.

[6] H. J. Sussmann, *Orbits of families of vector fields and integrability of systems with singularities*, Bull. Am. Math. Soc. 79 (1973), pp. 197–199.

P. Brunovský

On the Structure of Optimal Feedback Systems

In: Proceedings of the International Congress of Mathematicians Helsinki, (1978), 841–846.

Proceedings of the International Congress of Mathematicians
Helsinki, 1978

On the Structure of Optimal Feedback Systems

Pavol Brunovský

The basic optimal control problem is given by a system

$$\dot{x} = f(x, u),\ x \in R^n,\ u \in R^m, \tag{1}$$

a control domain

$$U \subset R^m \tag{2}$$

a performance index

$$J(u) = \int_0^T f^0(x, u)\, dt\ \ (f^0 \colon R^n \times R^m \to R), \tag{3}$$

initial and target states x_0, x_1 respectively. By an admissible control we understand a piecewise continuous function, defined on some interval of the real line with values in U. Under suitable regularity conditions on f^0, f every admissible control $u \colon [0, T] \to U$ when substituted into (1) defines a unique solution $x(t, u)$ starting at x_0 for $t=0$ (called the response of u). Substituting the control and its response into (3) for u, x respectively, gives a real value to J. One is interested in finding and studying the properties of the optimal control which steers the system from x_0 to x_1 (i.e. its response $x(t)$ called the optimal trajectory satisfies $x(T)=x_1$) for some $T>0$ and minimizes the performance index J.

From the very beginning of the optimal control theory one of the approaches to study this problem has been to imbed it in a family of problems with a varying initial state x_0. This approach is based on the simple observation (frequently called Bellman's optimality principle) that if u is an optimal control on $[0, T]$, then its restriction to any interval $[t_0, T]$, $t_0 \geqslant 0$, is an optimal control for the initial state $x(t_0, u)$. If for each initial state x in some region G the optimal control u_x (and, consequently, its response ξ_x starting at x) is unique, from the optimality principle

we obtain immediately that the optimal control can be expressed, independently of the initial state, as a function of the present state of the system, i.e. there exists a function $v: G \to U$ such that $u_x(t) = v(\xi_x(t))$ for $x \in G$. Therefore the optimal trajectories satisfy in G the differential equation

$$\dot{x} = f(x, v(x)). \tag{4}$$

Let us note that in many applications the ultimate goal of solving the optimal control problem is to find the function v, which is called the closed-loop optimal control, the optimal feedback law or the synthesis of optimal control.

Formally, one can consider (4) as an equation for optimal trajectories. In order to utilize it, it is important to know something about the properties of the function v. For example, for the classical existence and uniqueness theory of ordinary differential equations it would be useful if v were continuous. However, simple examples in which v can be constructed explicitly (cf. [**1**, Chapter III] or [**11**, Chapter 2]) show that due to unilateral constraints, which are typical for the optimal control theory, v is frequently discontinuous.

A deeper reason for studying the structure of v is the problem of sufficiency of the variational necessary conditions of optimality, in particular of the Pontrjagin maximum principle (PMP). Assume that for every initial state $x \in G$ there exists a unique control steering the system from x to x_1 and satisfying PMP, thus being the unique candidate for the optimal control. If we define $v(x) = u_x(0)$, we may ask whether u is the closed-loop optimal control, i.e. whether (4) yields optimal trajectories (and only optimal trajectories) as its solutions. As it is shown in [**1**], [**2**] this problem is closely connected with the problem of the sufficiency on the dynamic programming equation (which corresponds to the Hamilton–Jacobi equation of the classical calculus of variations).

When trying to resolve this question one is again confronted with the problem of the regularity of the behaviour of v. Bolt'anski observed that one can work also with a discontinuous synthesis, provided its set of discontinuities is sufficiently regular. This led him to introduce the concept of regular synthesis for the time-optimal control problem ($f^0 = 1$) (cf. [**1**], [**2**]). By a regular synthesis for the time-optimal control problem in a region G we understand a pair $(\mathscr{S}, v)$, where $\mathscr{S}$ is a locally finite partition of G into C^1 connected submanifolds of G (called cells), v is a function $G \to U$ satisfying the following conditions:

A. The set $\bar{G}'$ (where G' is the union of the cells of dimension $< n$) admits a stratification in G. (By a stratification $\mathscr{P}$ of a subset H of G we understand a locally finite partition of H into C^1 connected submanifolds of G (called strata) such that $P \cap \bar{Q} \neq \emptyset$ implies $P \subset \bar{Q}$ and $\dim P < \dim Q$ for any $P, Q \in \mathscr{P}$, $P \neq Q$.)

B. The function v is C^1 on each $S \in \mathscr{S}$ and can be extended to a C^1 function in some neighbourhood of S. The cells of $\mathscr{S}$ are of type I and type II. If S is

of type I, then $f(x, v(x)) \in T_x S$ (the tangent space of S at x) for every $x \in S$ and there is a uniquely defined cell $\pi(S)$ such that every solution of (4) starting at any point $x \in S$ enters $\pi(S)$ transversally for some $\tau > 0$ (after staying in S on $(0, \tau)$) which is a continuous function of x. If S is of type II then $f(x, v(x)) \notin T_x S$ for all $x \in S$ and there is a unique cell $\Sigma(S)$ of type I such that v is C^1 on $S \cup \Sigma(S)$ and every solution of (4) starting in S lies in $\Sigma(S)$ for sufficiently small positive times.

C. Every trajectory $x(t)$ of (4) starting at some point $x \in G$ (which is by B uniquely defined until it stays in G) eventually reaches x_1 in finite time $T(x) \geqslant 0$ passing through a finite number of cells only and together with the control $u(t) = v(x(t))$ satisfies PMP.

D. $T(x)$ is continuous in G. Let us note that this definition differs somewhat from Bolt'anski's one as well as from that of [3]. (For details, cf. [3] and the forthcoming Erratum to [3].)

In [2] (cf. also [1]) Bolt'anski proved that if $(\mathcal{S}, v)$ is a regular synthesis, then v is the closed-loop optimal control in the following sense:

The trajectory ξ_x (in the Carathéodory sense) on $[0, T(x)]$ of equation (4) starting at $x \in G$ is the optimal trajectory and $u_x(t) = v(\xi_x(t))$ is the optimal control.

Virtually in all the simple examples in which it has been possible to construct the synthesis explicitly, the latter has satisfied the conditions of regularity. However, except for some studies of the local structure of v near x_1 (cf. e.g. [14]) no attempt has been made to prove that a more general class of problems would globally admit a regular synthesis. Such a result has been made possible by Hironaka's theory of subanalytic sets [7], [9], [10]. It concerns linear control systems

$$\dot{x} = Ax + Bu \tag{5}$$

with

$$U = \text{co}\,\{w_1, \ldots, w_p\} \tag{6}$$

being a convex polytope. Such a problem is called normal if for every $i \neq j$, k,

$$\det\left(b_k(w_i - w_j), Ab_k(w_i - w_j), \ldots, A^{n-1}b_k(w_i - w_j)\right) \neq 0,$$

where $B = (b_1, \ldots, b_m)$. Let us note that normality is a generic property (cf. [11, Chapter 2, Theorem 11]).

THEOREM 1 [3]. *Assume that the control system defined by* (5), (6) *is normal and that* U *contains* 0 *in its interior. Then the time-optimal control problem with the target point* $x_1 = 0$ *admits a regular synthesis in the domain* G *of points that can be steered to* 0.

As mentioned above, the proof of this theorem makes use of the theory of subanalytic sets. A subset M of an analytic manifold is called subanalytic if it can be locally (in A) expressed as a finite union of sets of type $f(Y) \setminus g(Z)$, where Y, Z are analytic manifolds and f, g are analytic proper. By the central theorem of the

theory of subanalytic sets, every subanalytic subset of A admits an analytic stratification, the strata of which are subanalytic (cf. also [13]).

The cells of the synthesis are obtained by an inductive construction. The sets of continuity of v are shown to be subanalytic and the synthesis cells are obtained by a sequence of partitions of these sets into connected analytic submanifolds. In addition to the standard theory of subanalytic sets one needs the following

LEMMA. *Let M be a subanalytic subset of an analytic manifold A and let $X_1, \ldots, X_r$ be analytic vector fields on A. Then M admits a locally finite partition $\mathscr{P}$ into connected analytic submanifolds of A, which are subanalytic in A, such that for every $P \in \mathscr{P}$ and $i=1, \ldots, r$, X_i is either everywhere or nowhere tangent to P.*

This lemma, an improved version of which has been proved by Sussmann, appears to be crucial also for other application of the theory of subanalytic sets in control theory (cf. [12]). From this theorem it immediately follows that the minimum steering time to x_1, $T(x)$, is analytic in G everywhere except for a stratified set (G') of dimension $n-1$ (=maximal dimension of the strata).

If one tries to extend the concept of regular synthesis to problems where PMP yields controls with corners which are not jumps (like time optimal control problems with control domains having piecewise analytic curvilinear boundaries, or linear-quadratic problems with linear constraints), one immediately sees that the transversality assumptions as well as the C^1 extendability of v to the neighbourhood of S in B cannot be required. Instead, one has to assume their consequences, namely that the time $\tau(x)$ at which $\xi_x(t)$ enters $\pi(S)$ for S of type I and $\pi(\Sigma(S))$ for S of type II, the trajectory $\xi_x(t)$ and the control $u_x(t)$ are C^1 functions of x, t for $x \in S$, $t \in [0, \tau(x))$ and can be extended to C^1 functions of x, t for $t \geqslant \tau(x)$ close to $\tau(x)$. With this difference, the definition of regular synthesis can be literally extended to control problems with other performance indices (T in D replaced by J, the performance index). Bolt'anski's proof can be extended easily to yield an extension of his sufficiency theorem to general performance indices.

Employing essentially the same induction techniques as in the linear time-optimal problem case, one can prove an abstract existence theorem. However, due to the lack of transversality mentioned above, in order to obtain the C^1 dependence of the required quantities one has to construct auxiliary partitions in the product space of the state and adjoint space. By suitable partitions one can achieve that the product flow of the system and its adjoint enters the cells in the product space transversally, thus yielding analyticity of the required quantities.

Because of lack of space we desist from introducing this theorem, which has a rather cumbersome formulation. This is due to technical assumptions, which are needed for the extendability of the solutions of certain vector fields to sufficiently long intervals. Rather we note that the most serious requirements (in addition to analyticity, of course) for a system to admit a regular synthesis in some region G are the following ones:

1. For every initial state $x \in G$, there has to be a unique control u_x satisfying PMP which steers the system from x to x_1.

2. The number of switchings (which are roughly speaking the points of non-analyticity) of the controls u_x has to be locally uniformly bounded.

The first requirement makes the range of applications of such a result rather limited. Indeed, although singular controls (not minimizing the Hamiltonian strictly), which are quite typical for nonlinear control problems, are not excluded in principle, when they appear the first requirement is usually not satisfied. On the other hand the second requirement, the validity of which is difficult to prove for more general classes of systems, is virtually always satisfied in particular problems.

The following theorem concerns a model class of problems in which these difficulties can be overcome—linear-quadratic optimal control problems with linear constraints.

THEOREM 2. *Consider the optimal control problem*

$$\dot{x} = Ax + Bu,$$

$$J = \int_0^T [x^* Q x + u^* R u]\, dt \quad (R > 0,\ Q \geqslant 0),$$

$$U = \{u \in R^m \mid \langle l_j, u \rangle \leqslant m_j,\ j = 1, \ldots, p\},$$

$x(T) = 0$, *T fixed, and assume that this system is normal. Then the problem admits a regular synthesis.*

The normality assumption here consists in the non-vanishing of certain polynomials involving the entries of A, B, Q, l_j, m_j, as in the case of the linear time-optimal control problem it is a generic property.

Of course, this theorem has a similar impact on the regularity of the minimal value of the performance index as Theorem 1 had on the regularity of the minimal steering time.

Let us note that neither Theorem 1 nor Theorem 2 contribute anything to sufficient conditions of optimality (the sufficiency of PMP in both cases can be proved by other, simpler means). Their value lies rather in the insight they give into the structure of the closed-loop optimal control.

Finally let us note that in Bolt'anski's sufficiency results one understands the solutions in the classical Carathéodory sense. However, it has been demonstrated by several authors in the fifties that this concept is inadequate in the case of equations with discontinuities in the dependent variable. Because of the discontinuity of v this is the case for equation (4) in many control problems. Several concepts of solutions for such equations have been proposed, the most elaborate being that of Filippov [6]. Therefore it is natural to ask whether the optimal trajectories (which are the usual solutions of (4)) coincide with the Filippov trajectories or not. This problem is related to the problem of stability of the behaviour of the solutions of (4) with respect to perturbations (cf. [8], [4]). Using a slight improvement of Theorem 1 this question can be answered positively for the linear time-optimal control problem

with $\dim u=1$ (cf. [3], [5]). However, the results of [4], where the problem is completely solved for the two-dimensional linear time-optimal control problem, show that there is a non-exceptional class of problems for which the optimal trajectories do not coincide with the Filippov trajectories of (4).

The author is indebted to H. Sussmann whose comments on [3] have been of great value for the present paper.

References

1. V. G. Bolt'anski, *Mathematical methods of optimal control,* "Nauka", Moscow, 1969.

2. *Sufficient conditions of optimality and the foundations of the methods of dynamic programming.* Izv. Akad. Nauk. SSSR, Ser. Mat. **28** (1968), 481—514.

3. P. Brunovský, *Every normal linear system has a regular time-optimal synthesis,* Math. Slovaca **28** (1978), 81—100.

4. *The closed-loop time-optimal control* I, II, SIAM J. Control **12** (1974), 624—634; **14** (1976), 156—162.

5. P. Brunovský, and S. Mirica, *Classical and Filippov solutions of the differential equations defined by feedback control,* Rev. Roumaine Math. Pures Appl. **20** (1975), 873—883.

6. A. F. Filippov, *Differential equations with a discontinuous right-hand side,* Mat. Sb. **51** (1960), 99—128.

7. R. M. Hardt, *Stratification of real analytic mappings and images.* Invent. Math. **28** (1975), 193—208.

8. H. Hermes, *Discontinuous vector fields and feedback control,* Differential Equations and Dynamic Systems, J. K. Hale and J. P. La Salle, editors, Academic Press, New York, 1967. pp. 155—165.

9. H. Hironaka, *Introduction aux ensembles sous-analytiques.* Asterisque **7—8** (1973), 13—20.

10. *Subanalytic sets,* Number Theory, Algebraic geometry and Commutative Algebra, in Honour of Y. Akizuki, Kinokuniya, Tokyo, 1973, pp. 453—493.

11. E. B. Lee and L. Markus, *Foundations of optimal control theory,* Wiley, New York and London, 1967.

12. H. J. Sussmann, *Subanalytic sets and feedback control,* J. Differential Equations (to appear).

13. *Analytic stratifications and control theory,* these PROCEEDINGS.

14. D. S. Yeung, *Time-optimal feedback control,* J. Optimal Theory Appl. **21** (1977), 71—82.

INSTITUTE OF APPLIED MATHEMATICS, COMENIUS UNIVERSITY
MLYNSKA DOLINA
(81631) BRATISLAVA, CZECHOSLOVAKIA

P. Brunovský, J. Komorník

The matrix Riccati equation and the noncontrollabe linear-quadratic problem with terminal constraints

SIAM Journal on Control and Optimization, Vol. 21, No. 2 (1983), 280-288.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0008 $01.25/0

THE MATRIX RICCATI EQUATION AND THE NONCONTROLLABLE LINEAR-QUADRATIC PROBLEM WITH TERMINAL CONSTRAINTS*

PAVOL BRUNOVSKÝ† AND JOZEF KOMORNÍK‡

Abstract. It is proved that each positive semidefinite symmetric solution of the matrix Riccati equation corresponds to an optimal control problem with suitable terminal cost and constraints. The approximation scheme for the computation and characterization of the optimal cost and optimal controls of the problem with terminal constraints is extended to the noncontrollable case.

Key words. matrix Riccati, linear-quadratic, terminal constraints

Introduction. Consider the linear-quadratic optimal control problem on the interval $[s, T]$, $t_0 \leqq S \leqq T$, given by the equation

$$\dot{x} = A(t)x + B(t)u \tag{1}$$

($x \in R^n$, $u \in R^r$), the initial state

$$x(s) = y, \tag{2}$$

the cost function

$$C_s^T(y, u) = \int_s^T c(t, x, u)\, dt + x'(T)Rx(T) \tag{3}$$

with $c(t, x, u) = x'Q(t)x + u'M(t)u$ and the terminal constraint

$$Dx(T) = 0, \tag{4}$$

D being $q \times n$, $q \leqq n$, with full rank, A, B, Q, M being continuous, Q, M symmetric, $Q \geqq 0$, $M > 0$ on $[t_0, T]$, $R \geqq 0$ symmetric.

Under the condition that the system with output $\xi = Dx$ is output controllable on $[s, T]$ for each $t_0 \leqq s < T$, we have shown in [1] that the minimal cost for this problem can be expressed by a solution of the corresponding matrix Riccati equation

$$\dot{W} + A'W + WA + Q - W'BM^{-1}B'W = 0 \tag{5}$$

(cf. also [2]) on $[t_0, T)$ that blows up for $t \nearrow T$. We have characterized this solution as a limit for $m \to \infty$ of solutions of (5) expressing the optimal cost of the corresponding unconstrained problem with cost

$$C_{s,m}^T(y, u) = C_s^T(y, u) + m\|Dx(T)\|^2 \tag{6}$$

containing a term penalizing the deviation of the response of u from the terminal subspace. Also, we have shown that the optimal control and optimal trajectory for the problem (1)–(4) are limits for $m \to \infty$ of those for the problems (1)–(3), (6).

This result can be put into an interesting context with the ideas of [3]. By associating with (5) a flow on the Grassmann manifold $GR(n)$ of n-dimensional subspaces of R^{2n}, we can prove an inverse theorem on the solutions of (5) (§ 3) and extend our results from [1] to noncontrollable problems and problems with constraints at several points (§ 4). In § 5 we show that the techniques of § 4 can be used to deal with the infinite interval problem in case the finiteness of cost is not assumed for all

* Received by the editors March 11, 1981, and in revised form October 7, 1981.
† Institute of Applied Mathematics, Comenius University, 84215 Bratislava, Czechoslovakia.
‡ Department of Probability and Statistics, Comenius University, 84215 Bratislava, Czechoslovakia.

points. Section 2 contains a summary of the items of [3] that are important for this paper.

All the necessary material about the Riccati matrix equation and the unconstrained linear-quadratic problem is summarized in [1].

2. The associated flow on $GR(n)$. Denote

$$J = \begin{pmatrix} 0 & E_n \\ -E_n & 0 \end{pmatrix}$$

where E_n is the $n \times n$ unity matrix. For $z_i = (x_i, p_i) \in R^n \times R^n$, $i = 1, 2$, denote

$$\omega(z_1, z_2) = z_1' J z_2 = x_1' p_2 - x_2' p_1:$$

ω is a skew symmetric nondegenerate form on R^{2n}. An n-dimensional linear subspace L of R^{2n} is called Lagrangian, if the restriction of ω to L vanishes, i.e. $\omega(z_1, z_2) = 0$ as soon as $z_1, z_2 \in L$. We denote the set of Lagrangian subspaces of R^{2n} by $\mathscr{L}$.

A linear differential equation in R^{2n}

$$\dot{z} = H(t)z \tag{7}$$

is called Hamiltonian if ω is its integral, i.e. ω is constant along its solutions. This is equivalent to

$$H'J + JH = 0. \tag{8}$$

By π_x we denote the natural projection of $R^{2n} = R^n \times R^n$ onto its first factor, and we denote

$$\mathscr{L}_0 = \{L \in \mathscr{L} | \pi_x(L) = R^n\}.$$

We have $L \in \mathscr{L}_0$ if and only if there exists a symmetric $n \times n$ matrix W such that

$$L = \{(x, Wx) | x \in R^n\}.$$

The (time-dependent) flow of the equation (7) carries linear subspaces into linear subspaces of the same dimension and thus generates an associated (time-dependent) flow Φ on the Grassmann manifold $GR(n)$ of the subspaces of R^{2n} of dimension n. More precisely, if $L \in GR(n)$ and we denote by $\Phi_{t,s}(L)$ the linear subspace filled by the values at t of the solutions of (7) with values in L at time s, then there is a differential equation on $GR(n)$ such that $\Phi_{t,s}(L)$ is the value at time t of its solution having L as its value at time s. Since $GR(n)$ is compact, the solutions of this equation are defined for all $t \in R$. Since ω is an integral of (7), it is invariant under Φ, i.e. $L \in \mathscr{L}$ implies $\Phi_{t,s}(L) \in \mathscr{L}$ for all $t, s \in R$.

Consider now the flow Φ on $GR(n)$ associated with the differential equation

$$\dot{x} = Ax - BM^{-1}B'p, \qquad \dot{p} = -Qx - A'p \tag{9}$$

with A, B, M, Q coming from (1), (3). The matrix

$$H = \begin{pmatrix} A, & -BM^{-1}B' \\ -Q & -A \end{pmatrix}$$

obviously satisfies (8), which means that (9) is Hamiltonian. If $L \in \mathscr{L}$ and $L(t) = \Phi_{t,s}(L) \in \mathscr{L}_0$ for all $t \in I = (t_1, t_2)$ and some $s \in I$, then there exists a matrix function $W(t)$, $t \in (t_1, t_2)$, such that $L(t) = \{(x, W(t)x) | x \in R^n\}$. This matrix satisfies (5).

Note that although $\lim_{t \to t_0} W(t)$ may not exist for $t_0 = t_1$ or $t_0 = t_2$, $L(t) = \Phi_{t,s}(L)$ can always be extended beyond I to all R.

3. The inverse theorem. Let A, B, Q, M be as in (1) (3).

THEOREM 1. *Let $W(t)$ be a positive semidefinite solution of the matrix Riccati equation* (5) *on* $[t_0, T)$. *Then there exist a* $q \leqq n$, *a* $q \times n$ *matrix* D *and a positive semidefinite symmetric matrix* R *such that* $y'W(s)y$ *is the optimal cost for the problem* (1)–(4) *for* $t_0 \leqq s < T$.

Proof. Denote $L(t) = \{(x, W(t)x) | x \in R^n\}$. If $\lim_{t \to T-} W(t)$ exists then we take $D = 0$, $R = \lim_{t \to T-} W(t)$; the statement of the theorem in this case is standard [4].

If $\lim_{t \to T-} W(t)$ does not exist, then $L(T) = \lim_{t \to T-} L(t) \notin \mathscr{L}_0$. Let $q = \operatorname{codim} \pi_x(L(T)) > 0$.

There exist $n \times n$ matrices S_1, S_2 such that rank

$$(S_1, S_2) = n \quad \text{and} \quad L(T) = \{(x, p) | S_1 x + S_2 p = 0\}.$$

If $x \in \pi_x(L(T))$ then there exists a p such that $S_2 p = -S_1 x$, i.e. $S_1 x \in \operatorname{Range} S_2$. The condition rank $(S_1, S_2) = n$ means

$$\operatorname{Range} S_1 + \operatorname{Range} S_2 = R^n, \tag{10}$$

from which it follows codim Range $S_2 = q$. Consequently, there exists a $q \times n$ matrix with full rank N such that $y \in \operatorname{Range} S_2$ if and only if $Ny = 0$. From (10) it also follows that if we denote $D = NS_1$, then rank $D = \operatorname{rank} N = q$. Also, $x \in \pi_x(L(T))$ if and only if $Dx = 0$, i.e., $x \in \operatorname{Ker} D$.

Let K be any $n \times n$ matrix, the restriction of which to Range S_2 is a right inverse of S_2, i.e. we have $S_2 K S_2 = S_2$. Then, $(x, p) \in L(T)$ if and only if $x \in \operatorname{Ker} D$ and

$$p + KS_1 x \in \operatorname{Ker} S_2. \tag{11}$$

Denote $R_0 = -KS_1$.

Since $L(T) \in \mathscr{L}$, for any $p_1, p_2 \in \operatorname{Ker} S_2$, $x_1, x_2 \in \operatorname{Ker} D$ we have

$$(R_0 x_1 + p_1)' x_2 - (R_0 x_2 + p_2)' x_1 = 0. \tag{12}$$

Choosing $x_1 = 0$ and using (12) we obtain $p'x = 0$ for any $p \in \operatorname{Ker} S_2$, $x \in \operatorname{Ker} D$. However, $p'x = 0$ for all $x \in \operatorname{Ker} D$ is equivalent to $p \in \operatorname{Range} D'$, so $\operatorname{Ker} S_2 \subset \operatorname{Range} D'$. Since rank $D = q = \operatorname{codim} \operatorname{Range} S_2 = \dim \operatorname{Ker} S_2$, we have

$$\operatorname{Ker} S_2 = \operatorname{Range} D'. \tag{13}$$

Choosing $p_1 = p_2 = 0$ in (12) we have $x_1' R_0' x_2 = x_1' R_0 x_2$ for all $x_1, x_2 \in \operatorname{Ker} D$. Also, if we take any $x \in \operatorname{Ker} D$, then $(x, R_0 x) \in L(T)$. Since $L(T) = \lim_{t \to T-} L(t)$ (in $GR(n)$), there exists a sequence of points $t_i \nearrow T$, $(x_i, p_i) = (x_i, W(t_i)x_i) \in L(t_i)$, $(x_i, p_i) \to (x, R_0 x)$. Since $W(t)$ is positive semidefinite, for each $t < T$, we have $x' R_0 x = \lim_{i \to \infty} x_i' W(t_i) x_i \geqq 0$.

Denote $R_0^1 = PR_0$, $R_0^2 = (E - P)R_0$, where P is the orthogonal projection of R^n onto $\operatorname{Ker} D$. For any $x_1, x_2 \in \operatorname{Ker} D$ we have

$$x_1' R_0 x_2 = x_1' R_0^1 x_2$$

and, consequently,

$$x_1' R_0^1 x_1 \geqq 0, \qquad x_1' R_0^1 x_2 = x_2' R_0^1 x_1.$$

Thus, the restriction of R_0^1 to $\operatorname{Ker} D$ is symmetric and positive semidefinite. Obviously, we can find an R symmetric and positive semidefinite on all R^n such that

$$R|_{\operatorname{Ker} D} = R_0^1|_{\operatorname{Ker} D}. \tag{14}$$

By (13), for $x \in \operatorname{Ker} D$, (11) is equivalent to

$$p - R_0 x = p - R_0^1 x - R_0^2 x \in \operatorname{Range} D'.$$

Since $R_0^2 x$ is orthogonal to $\operatorname{Ker} D$, we have $R_0^2 x \in \operatorname{Range} D'$, which means that (11) is equivalent to

$$p - Rx \in \operatorname{Range} D' \tag{15}$$

for all $x \in \operatorname{Ker} D$. By [5], [2], (14) is the transversality condition for the solution of the adjoint equation of the problem (1)–(4). Since $R \geqq 0$, $M(t) > 0$ and $Q(t) \geqq 0$ for $t \in [t_0, T]$, if $(x(t), p(t))$ is a solution of (9) with $Dx(T) = 0$ and $p(T)$ satisfying (15), then $x(t)$, $t \in [s, T]$ is an optimal trajectory for the problem (1), (3), (4) with initial state $x(s)$, the corresponding optimal control being generated by the feedback law

$$u(t) = -M^{-1}(t)B(t)p(t) = -M^{-1}(t)B(t)W(t)x(t) \tag{16}$$

for $s \leqq t < T$. Since $\pi_x(L(S)) = R^n$, the points $x(s)$ obtained in this way for all possible choices of $x(T)$ and $p(T)$ fill up all R^n.

We have

$$\begin{aligned} x'(s)W(s)x(s) = p(s)x(s) &= -\int_s^T \frac{d}{dt}(p(t)x(t))\,dt + p'(T)x(T) \\ &= x'(T)Rx(T) - \int_s^T [\dot{p}'(t)x(t) + p'(t)\dot{x}(t)]\,dt \\ &= x'(T)Rx(T) + \int_s^T [x'(t)Q(t)x(t) + u'(t)M(t)u(t)]\,dt. \end{aligned}$$

Since $x(t)$, $u(t)$ are the optimal trajectory and control, respectively, this completes the proof.

4. The noncontrollable problem. In this section we consider the problem (1)–(4), but unlike in [1], [2], we shall not assume that the system (1) with output $\xi = Dx$ is output controllable. It is obvious that the set of points that can be controlled to the terminal set $Dx(T) = 0$ on $[s, T]$ is a linear subspace of R^n, but for a nonautonomous problem it is moving with s in general, and it is not entirely obvious how to characterize it.

The following theorem gives two characterizations of this subspace—one in terms of the flow on $GR(n)$, the other in terms of the approximation scheme of [1]. Also, it shows that for this approximation scheme to work, the output controllability assumption is not essential.

As in [1], we denote by $\mathcal{U}_s^T(y)$ the set of controls steering the system from the point y to the terminal set (4) on $[s, T]$ and by W_m the solution of (5) satisfying the terminal condition $W_m(T) = R + mD'D$. Note that $y'W_m(s)y$ is the minimal value of the cost for the unconstrained problem (1)–(3), (6). The optimal control $u_m(t)$ for this problem is given by the optimal feedback law

$$u = -M^{-1}(t)B'(t)W_m(t)x, \tag{17}$$

i.e., we have $u_m(t) = -M^{-1}(t)B'(t)W_m(t)x_m(t)$, where $x_m(t)$ is the solution of the equation

$$\dot{x} = (A - BM^{-1}B'W_m)x$$

with $x_m(s) = y$.

Denote

$$U(s)=\{y\,|\,\mathcal{U}_s^T(y)\neq\emptyset\},$$

$$V(s)=\{y\,|\,\limsup_{m\to\infty} y'W_m(s)y<\infty\},$$

$$L(s)=\Phi_{s,T}(\{(x,p)|Dx=0,\ p-Rx\in\text{Range } D'\}).$$

THEOREM 2. *For all* $s\in[t_0,T)$,

$$U(s)=V(s)=\pi_x(L(s)). \tag{18}$$

For $y\in U(s)$, *the optimal control* $u_0(t)$ *for the problem* (1)–(4) *is given by*

$$u_0(t)=\lim_{m\to\infty}u_m(t)=-\lim_{m\to\infty}M^{-1}(t)B'(t)W_m(t)x_m(t),$$

and the optimal value of the cost is given by

$$\min_{u\in\mathcal{U}_s^T(t)} C_s^T(y,u)=C_s^T(y,u_0)=\lim_{m\to\infty}y'W_m(s)y. \tag{19}$$

Proof. First, we prove $V(s)\subset U(s)$. From (9) we obtain by simple calculation for any k, m, s fixed, $y=x_i(s)$ and $p_i(t)=W_i(t)x_i(t)$, $i=k,m$,

$$\begin{aligned}
&\int_s^T[(x_m(t)-x_k(t))'Q(t)(x_m(t)-x_k(t))+(u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))]\,dt\\
&\quad=-\int_s^T\frac{d}{dt}[(p_m(t)-p_k(t))'(x_m(t)-x_k(t))]\,dt\\
&\quad=-(p_m(T)-p_k(T))'(x_m(T)-x_k(T))\\
&\quad=-(x_m(T)-x_k(T))'(W_m(T)x_m(T)-W_k(T)x_k(T))\\
&\quad=-(x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T))\\
&\qquad-(x_m(T)-x_k(T))'(m-k)D'Dx_m(T)\\
&\quad=-(x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T))\\
&\qquad-(m-k)x_m'(T)D'Dx_m(T)+x_k'(T)(W_m(T)-W_k(T))x_m(T).
\end{aligned} \tag{20}$$

Using the invariance of ω, we have

$$\begin{aligned}
x_k'(T)(W_m(T)-W_k(T))x_m(T)&=p_m'(T)x_k(T)-p_k'(T)x_m(T)\\
&=p_m'(s)x_k(s)-p_k'(s)x_m(s)\\
&=y'(W_m(s)-W_k(s))y.
\end{aligned}$$

Denote $\delta_{k,m}(s)=y'(W_m(s)-W_k(s))y$.

$$\bar{\delta}_k(s)=\lim_{m\to\infty}\delta_{k,m}(s);$$

$$0\leqq\delta_{k,m}(s)\leqq\bar{\delta}_k(s)<\infty,\quad\lim_{k\to\infty}\bar{\delta}_k(s)=0\quad\text{for } k<m,\ y\in V(s). \tag{21}$$

From (20) it follows that

$$\begin{aligned}\delta_{k,m}(s) = \int_s^T & [(x_m(t)-x_k(t))'Q(t)(x_m(t)-x_k(t)) \\ & + (u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))]\,dt \\ & + (x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T)) \\ & + (m-k)x_m'(T)D'Dx_m(T).\end{aligned} \tag{22}$$

Since all the right-hand side terms are nonnegative, we have

$$\begin{aligned}&(m-k)\|Dx_m(T)\|^2 = (m-k)x_m'(T)D'Dx_m(T) = \delta_{k,m}(s) \leqq \bar{\delta}_k(s), \\ &0 \leqq \|Dx_m(T)\|^2 \leqq \frac{1}{m-k}\bar{\delta}_k(s),\end{aligned} \tag{23}$$

and, by (21),

$$\lim_{m\to\infty} \|Dx_m(T)\| = 0.$$

Also, from (22) it follows that

$$\sup_m \int_s^T (u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))\,dt \leqq \bar{\delta}_k(s).$$

Since $M(t)$ is continuous and positive definite on $[s, T]$, it is uniformly positive definite on $[s, T]$. From this and (21) it follows that $\{u_m\}$ is a Cauchy sequence in $L_2(s, T)$ and therefore has a limit $u_0(t)$ in $L_2(s, T)$. From the representation of $x_m(t)$ by the variation of constant formula it follows immediately that $\{x_m\}$ converges uniformly to the response $x_0(t)$ of $u_0(t)$ satisfying $x_0(s) = y$.

By (23), we have

$$Dx_0(T) = 0.$$

This proves $V(s) \subset U(s)$ and also the second equality of (19). To prove the first equality (having as its consequence the optimality of u_0) we note that for each $u \in \mathcal{U}_s^T(y)$ we have

$$C_s^T(y, u) = C_{s,m}^T(y, u) \geqq \min C_{s,m}^T(y, u) = y'W_m(s)y.$$

This also proves $U(s) \subset V(s)$. To complete the proof of the theorem it remains to prove the second equality of (18).

If $y \in \pi_x(L(s))$ then there exists a solution $(x(t), p(t))$ of (9) with $Dx(T) = 0$ such that $x(s) = y$. The function $x(t)$ is a response of the control $u(t) = -M^{-1}(t)B'(t)p(t)$ which means $u \in \mathcal{U}_s^T(y)$. Consequently, $\mathcal{U}_s^T(y) \neq \emptyset$ and $y \in U(s)$.

On the other hand, if $y \in U(s)$, then by [5], there exists an optimal control u_0 in $\mathcal{U}_s^T(y)$, the response $x_0(t)$ of which, together with a suitable function $p(t)$, satisfies (9). In addition, $p(T)$ satisfies the transversality condition (15). This proves $y \in \pi_x(L(s))$.

Remark 1. Since $u_m \to u_0$ in $L_2(s, T)$ we have

$$\lim_{m\to\infty} C_s^T(y, u_m) = C_s^T(y, u_0). \tag{24}$$

On the other hand, we have

$$\begin{aligned}C_s^T(y, u_0) &= \lim_{m\to\infty} y'W_m(s)y = \lim_{m\to\infty} C_{s,m}^T(y, u_m) \\ &= \lim_{m\to\infty} C_s^T(y, u_m) + m\|Dx_m(T)\|^2.\end{aligned} \tag{25}$$

From (24), (25) we obtain

$$\lim_{m\to\infty} m\|Dx_m(T)\|^2 = 0,$$

or

$$\|Dx_m(T)\|^2 = o(m^{-1/2}).$$

This gives an estimate for the deviation of the endpoint of the optimal trajectory of the approximate unconstrained problem from the terminal set.

Remark 2. From $\pi_x(L(s)) = U(s)$ it follows that the dimension of $\pi_x(L(s))$ cannot decrease with s decreasing. From [6] it follows that for A, B analytic it is constant for $s < T$ and equal to the dimension of the space $\operatorname{Ker} D + C$, where $C = \operatorname{span}\{b_i(T), (\mathscr{A}b_i)(T), \cdots, (\mathscr{A}^{n-1}b_i)(T) | i = 1, \cdots, n\}$, where b_i are the column vectors of B and $\mathscr{A}f(t) = f(t) - Af(t)$ for a differentiable function f on $[t_0, T]$.

Theorem 2 allows us to deal with the problem (1)–(3) with additional constraints and costs at intermediate points of the interval. We shall restrict ourselves to the case of one intermediate point, the extension to the case of a higher number of points being straightforward.

Let $T_1 \in (t_0, T)$, $q_1 \leqq n$ and let $R_1 \geqq 0$, D_1 be $n \times n$ symmetric and $q_1 \times n$ with full rank, respectively. Consider the problem given by the system (1), the initial point (2), the cost function

$$\tilde{C}_s^T(y, u) = C_s^T(y, u) + x'(T_1)R_1x(T_1), \tag{26}$$

the constraints (4) and

$$D_1x(T_1) = 0. \tag{27}$$

Of course, for $s \in (T_1, T]$ the problem coincides with the problem (1)–(4).

Let $U(t)$, $W_m(t)$ be defined as in Theorem 2. It is obvious that the optimal control for the problem (1), (2), (26), (4), (27) for $s = T_1$ will be a concatenation of the optimal control on $[s, T_1]$ for the problem (1), (2), the cost function

$$\int_s^{T_1} c(t, x, u)\, dt + x'(T_1)R_1x(T_1) + \lim_{M\to\infty} x'(T_1)W_m(T_1)x(T_1)$$

and the linear constraint

$$x(T_1) \in U(T_1) \cap \operatorname{Ker} D_1,$$

and the optimal control for the problem (1), (3), (4), with initial point $x(t_1)$ on $[t_1, T]$.

5. The infinite interval. Consider the unconstrained problem (1), (3) with $R = 0$ and denote W^T the corresponding solution of (5), which is the solution satisfying $W^T(T) = 0$. For fixed s, y, denote u^T, x^T the optimal control and trajectory respectively. In [1], we have shown that $\lim_{T\to\infty} W^T(s)$ exists and represents the optimal cost for the infinite interval problem, provided for each s, y there exists a u such that $C_s^\infty(y, u) = \lim_{T\to\infty} C_s^T(y, u) < \infty$. Like Theorem 2, the following theorem deals with problems not satisfying this condition.

By $U^\infty(s)$ we denote the set of those $y \in R^n$ for which there is a control u on $[s, \infty)$ such that $C_s^T(y, u) < \infty$. Further, we denote

$$V^\infty(s) = \left\{ y \,\middle|\, \lim_{T\to\infty} y'W^T(s)y < \infty \right\}.$$

By $L^2_M(s, \infty)$ we denote the space of functions $u : [s, \infty) \to R^r$ which are square integrable with weight $M(t)$, i.e., $\int_s^\infty u'(t)M(t)u(t)\,dt < \infty$. $L^2_M(s, \infty)$ is a Banach space.

THEOREM 3. *We have $U^\infty(s) = V^\infty(s)$ for every $s \geqq t_0$. For $y \in U^\infty(s)$ we have*

$$\min_u C_s^\infty(y, u) = \lim_{T\to\infty} y'W^T(s)y.$$

The optimal control $u^\infty(t)$ and trajectory $x^\infty(t)$ are given by

$$u^\infty(t) = \lim_{T\to\infty} u^T(t) \qquad (\text{in } L^2_M(s, \infty)), \tag{28}$$

$$x^\infty(t) = \lim_{T\to\infty} x^T(t) \tag{29}$$

(uniformly on each finite interval).

Let us note that in (28), (29) we understand u^T, x^T to be extended to $[s, \infty)$ by having value 0 for $t > T$.

Proof. Let $y \in V^\infty(s)$, $T_2 = T_1 \geqq s$. Denote $W^{T_i} = W_i$, $x^{T_i} = x_i$, $u^{T_i} = u_i$, $i = 1, 2$. By computations similar to those leading to (20) we obtain

$$\begin{aligned} y'(W_1(s) - W_2(s))y &= \int_s^T [(x_1(t) - x_2(t))'Q(t)(x_1(t) - x_2(t)) \\ &\qquad + (u_1(t) - u_2(t))'M(t)(u_1(t) - u_2(t))]\,dt \\ &\quad + (x_1(T_1) - x_2(T_1))W_1(T_1)(x_1(T_1) - x_2(T_1)) \\ &\quad + x_2(T_1)(W_2(T_1) - W_1(T_1))x_2(T_1) \\ &= \int_s^{T_1} [(x_1(t) - x_2(t))'Q(t)(x_1(t) - x_2(t)) + (u_1(t) \\ &\qquad - u_2(t))'M(t)(u_1(t) - u_2(t))]\,dt \\ &\quad + \int_{T_1}^{T_2} [x_2(t)'Q(t)x_2(t) + u_2'(t)M(t)u_2(2)]\,dt \\ &\geqq \int_s^{T_1} (u_1(t) - u_2(t))'M(t)(u_1(t) - u_2(t))\,dt. \end{aligned} \tag{30}$$

From the estimate (30) it follows that the family of functions $\{u_T | T \geqq s\}$ is a Cauchy family in $L^2_M(s, \infty)$. Since $L^2_M(s, \infty)$ is complete, it has a limit $u \in L^2_M(s, \infty)$. From the variation of constants formula it follows immediately that the response x^∞ of u^∞ is a pointwise limit of the functions x^T, the convergence being uniform on each finite subinterval of $[s, \infty)$.

For every fixed $T_0 \geqq s$ we have

$$C_s^{T_0}(y, u^\infty) = \lim_{T\to\infty} C_s^{T_0}(y, u^T) = \lim_{T\to\infty} C_s^T(y, u^T) = \lim_{T\to\infty} y'W^T(s)y,$$

from which it follows that $C_s^\infty(y, u^\infty)$ is finite and, thus, that $V^\infty(s) \subset U^\infty(s)$. On the other hand, we have for any control u,

$$C_s^{T_0}(y, u) \geqq y'W^{T_0}(s)y. \tag{31}$$

From (30), (31) it follows that

$$C_s^\infty(y, u) \geqq C_s^\infty(y, u^\infty) = \lim_{T\to\infty} y'W^T(s)y,$$

which implies that u^∞ is optimal.

The inclusion $U^{\infty}(s) \subset V^{\infty}(s)$ follows immediately from (31).

Note added in proof. There is an overlap of our § 4 and the paper of G. Chen and W. Mills, *Finite elements and terminal penalization for quadratic cost optimal control problems governed by ordinary differential equations*, this Journal, 19 (1981), pp. 744–764. In particular, the essential part of Theorem 3 or our paper is contained in Theorem 2.2 of the quoted paper.

REFERENCES

[1] P. BRUNOVSKÝ AND J. KOMORNÍK, *The Riccati equation solution of the linear-quadratic problem with constrained terminal state*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 398–402.

[2] B. FRIEDLAND, *On solutions of the Riccati equation in optimization problems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 303–304.

[3] R. HERMANN, *Cartanian Geometry, Nonlinear Waves and Control Theory*, Part A, Math. Science Press, Brookline, MA, 1979.

[4] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika (Prague), 9 (1973), pp. 42–61.

[5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[6] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of non-linear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

P. Brunovský

On one-parameter families of diffeomorphisms. II. Generic branching in higher dimensions

Comment. Math. Univ. Carolinae 12 (1971), 765–784.

Commentationes Mathematicae Universitatis Carolinae

12,4 (1971)

ON ONE-PARAMETER FAMILIES OF DIFFEOMORPHISMS II: GENERIC BRANCHING IN HIGHER DIMENSIONS

Pavol BRUNOVSKÝ, Bratislava

§ 1

In [1], we have studied the generic nature of the loci of periodic points of a diffeomorphism of a finite dimensional manifold M, depending on a parameter with values in a one dimensional manifold P, in $P \times M$. A part of the results (those concerning the branching of periodic points), we have proved for two dimensional M only. It is the purpose of this paper to extend these results for M of arbitrary finite dimension.

Since this paper is a direct continuation of [1], we shall frequently refer to [1] for results of technical character as well as techniques of proof. Nevertheless, for the sake of the reader's convenience, we re-introduce those concepts and results of [1] which are necessary for the understanding of this paper, in the rest of this section. The main results of this paper and their proofs are given in § 3. § 2 has an auxiliary character; it establishes certain generic properties of maps of an interval into the

---

AMS: Primary 54H20
Secondary 57D50

Ref. Ž. 7.977.3

set of matrices.

Denote $\mathcal{F}$ the space of C^r mappings $(1 < r \leq \infty)$ [x] $f : P \times M \to M$, where P, M are C^r second countable manifolds of dimension 1, $m < \infty$ respectively, such that for every $p \in P$ the map $f_p : M \to M$, given by $f_p(m) = f(p, m)$ is a diffeomorphism, endowed with the C^r Whitney topology.

Let us note that, although this topology is not metrizable, it has the property that a residual set in $\mathcal{F}$ (i.e. a countable intersection of open dense sets) is dense in $\mathcal{F}$ (this can be proved similarly as the analogous statement for vector fields is proved in [2], using the openness of $\mathcal{F}$ in the set of all C^r mappings $P \times M \to M$).

Denote by $Z_k = Z_k(f)$ the set of k-periodic points of f, i.e. $Z_k(f) = \{(p, m) \mid f_p^k(m) = m$, $f_p^j(m) \neq m$ for $0 < j < k\}$. In [1, Theorem 1] a residual subset $\mathcal{F}_1$ of $\mathcal{F}$ was defined and it was shown that for every $f \in \mathcal{F}_1$, Z_k are one dimensional submanifolds of $P \times M$ (Z_1 being closed) and, if an eigenvalue of $df_p^k(m)$ at some point $(p, m) \in Z_k$ is 1 (we denote the set of such points by X_k), then it meets the unit circle S in the complex plain transversally at (p, m) (in the sense of Remark 3) and the remaining eigenvalues of $df_p^k(m)$ do not lie on S. Also, it was shown that the subset $\mathcal{F}_k$ of maps from $\mathcal{F}$, having the

---

x) In [1] we have assumed $1 < r < \infty$, but Theorems 1 - 4 of [1] are trivially true for the C^∞ case.

above properties for $1 \leq k \leq n$, is open dense in $\mathcal{F}$.

§ 2

Denote by $\mathcal{A}$ the set of all $n \times n$ matrices with the differential structure induced by its natural identification with R^{n^2}. Further, denote by $\mathcal{A}_1$ the set of matrices having an eigenvalue of multiplicity ≥ 2 on S, $\mathcal{F}_{2l}$ the set of matrices having an l-th root of unity different from ± 1 as eigenvalue, $\mathcal{A}_2 = \bigcup_{l=3}^{\infty} \mathcal{A}_{2l}$.

Let I be a closed interval on R. Denote by Φ the space of all C^r mappings $I \longrightarrow \mathcal{A}$ endowed with the C^r uniform topology.

Proposition 1. Let $J \subset I$ be a closed interval, $J \subset int\ I$. Then, for every $l = 3, 4, \ldots$ the set $\Psi_l(J)$ of all $F \in \Phi$ such that $F(J) \cap (\mathcal{A}_1 \cup \mathcal{A}_2) = \emptyset$ is open dense in Φ.

Corollary 1. Given J as in Proposiion 1, the set $\Psi(J)$ of all $F \in \Phi$ such that $F(J) \cap (\mathcal{A}_1 \cap \mathcal{A}_2) = \emptyset$ is residual in Φ.

For the proof of Proposition 1 we shall need to prove several lemmas.

Consider the sets $\tilde{\mathcal{A}}_1 = \{(A, \lambda_1, \lambda_2) \in \mathcal{A} \times R^2 \mid P_1(\lambda_1, \lambda_2) = P_2(\lambda_1, \lambda_2) = P_1'(\lambda_1, \lambda_2) = P_2'(\lambda_1, \lambda_2) = 0, \lambda_1^2 + \lambda_2^2 = 1\}$ and $\mathcal{A}_2(\lambda_{10}, \lambda_{20}) = \{(A, \lambda_1, \lambda_2) \mid P_1(\lambda_1, \lambda_2) = P_2(\lambda_1, \lambda_2) = 0, \lambda_1 = \lambda_{10}, \lambda_2 = \lambda_{20}\}$, where $P(\lambda_1) = P_1(Re\ \lambda, Im\ \lambda) + i P_2(Re\ \lambda, Im\ \lambda)$ is the characteristic polynomial of

$A, P_1' + i P_2' = P' = \frac{\partial P}{\partial \lambda}$.

Being defined by polynomial equalities, $\tilde{\mathcal{U}}_1$ and $\tilde{\mathcal{U}}_2(\lambda_{10}, \lambda_{20})$ are real algebraic varieties and the sets $\mathcal{U}_1$, $\mathcal{U}_{2\ell}$ are the projections of $\tilde{\mathcal{U}}_1$ and $\cup \tilde{\mathcal{U}}_2(\lambda_{10}, \lambda_{20})$ into $\mathcal{U}$ respectively, where the union is taken over all $\lambda_{10}, \lambda_{20}$ such that $(\lambda_{10} + i\lambda_{20})^\ell = 1$ and $\lambda_{20} \neq 0$.

By [3, splitting (b) of § 11)], $\tilde{\mathcal{U}}_1$ and $\tilde{\mathcal{U}}_2$ can be written as a finite disjoint union of submanifolds of strictly decreasing dimensions, $\tilde{\mathcal{U}}_1 = \dot{\bigcup}_{j=1}^{r} M_j$, $\tilde{\mathcal{U}}_2(\lambda_{10}, \lambda_{20}) = \dot{\bigcup}_{j=1}^{s} N_j$ such that $\dot{\bigcup}_{j=\varrho}^{r} M_j$, $\dot{\bigcup}_{j=\sigma}^{s} N_j$ is closed for all $0 < \varrho \leq r$, $0 < \sigma \leq s$.

Lemma 1. $\operatorname{codim} M_j \geq 4$ for all j .

For the proof of this lemma we need some more lemmas.

Lemma 2. For any $A \in \mathcal{U}$, the set of all matrices similar to A is an immersed submanifold of $\mathcal{U}$ of codimension $\geq n$.

Proof. Consider the group $GL(n)$, whose action ψ on $\mathcal{U}$ is given by $\psi(T,A) = T^{-1}AT$ for $T \in GL(n)$, $A \in \mathcal{U}$. The set of matrices similar to $\mathcal{U}$ is the orbit of A under this group action and, according to [4,2.2, Proposition 2], is an immersed submanifold of $\mathcal{U}$ of codimention equal to the dimension of the closed Lie subgroup $H = \{T \in GL(n) \mid \psi(T, A) = A\}$. It is easy to show that H is identical with the subset of $GL(n)$ of matrices that commute with A . It follows from [5,VIII, §2, Theorem 2] that H has the dimension $\geq n$, q.e.d.

Corollary 2. Denote by p the map $\mathcal{A} \to R^n$ assigning to every matrix from $\mathcal{A}$ the n-tuple of coefficients of its characteristic polynomial and $\tilde{p} : \tilde{\mathcal{A}} \to \to R^{n+2}$ as $\tilde{p} = p \times id$. Then, for any point $x \in R^{n+2}$, $p^{-1}(x)$ is a finite disjoint union of immersed submanifolds of $\tilde{\mathcal{A}}$ of codimension $\geq n$.

Denote by $V \subset R^{n+2}$ the set of points $(\alpha_1, \ldots, \alpha_n, \lambda_1, \lambda_2)$ such that $\lambda = \lambda_1 + i\lambda_2 \in S$ and is a root of the polynomial $P(\lambda) = \lambda^n + \alpha_1 \lambda^{n-1} + \ldots + \alpha_n$ of multiplicity ≥ 2. Obviously, $\tilde{p}(\tilde{\mathcal{A}}_1) = V$.

Lemma 3. The map $\tilde{p}|_{\tilde{\mathcal{A}}_1} : \tilde{\mathcal{A}}_1 \to V$ is open (in the topologies on $\tilde{\mathcal{A}}_1$, V induced by their imbedding into $\tilde{\mathcal{A}}$, R^{n+2} respectively).

Proof. Obviously, it suffices to prove that $p|_{\mathcal{A}_1} : \mathcal{A}_1 \to \hat{V}$, where $\hat{V}$ is the projection $(R^n \times R^2 \to \to R^n)$ of V into R^n, is open. That is, we have to prove that given a neighbourhood U of $A \in \mathcal{A}_1$, for any $P \in \hat{V}$ sufficiently close to $p(A)$, there is a $B \in U$ such that $p(B) = P$.

This statement is obvious if A has the real canonical form; its extension for A not in canonical form follows from $p(A) = p(T^{-1}AT)$ for $T \in GL(n)$.

Proof of Lemma 1. V is án algebraic variety in R^{n+2}, defined by the polynomial identities $P_1(\lambda_1, \lambda_2) = P_2(\lambda_1, \lambda_2) = P_1'(\lambda_1, \lambda_2) = P_2'(\lambda_1, \lambda_2) = \lambda_1^2 + \lambda_2^2 - 1 = 0$, where $P_1(\lambda_1, \lambda_2) = \mathrm{Re}\, P(\lambda_1 + i\lambda_2)$ etc. Therefore, it can be written as a finite disjoint union of submani-

folds of R^{n+2} of decreasing dimension, $V = \bigcup_{i=1}^{2} V_i$.

We prove $\dim V_1 \leq n-2$. To do this, we note that $\operatorname{codim} V_1 \geq \operatorname{rank}_x V$ for any $x \in V_1$ (cf. [3]), where $\operatorname{rank}_x V$ is the dimension of the linear space spanned by the differentials at x of the polynomials of the ideal associated with V . Since V_1 is open in V it suffices to prove that the set of those x for which $\operatorname{rank}_x V \geq 4$ is dense in V .

For $x \in V$, $x = (\alpha_1, \ldots, \alpha_n, \lambda_1, \lambda_2)$ we have

$$d P_1 = (\ldots, \lambda_1, 1, 0, 0) ,$$

$$(1) \quad d P_1' = (\ldots, 1, 0, \frac{\partial P_1'}{\partial \lambda_1}, \frac{\partial P_1'}{\partial \lambda_2}) ,$$

$$d P_2' = (\ldots, 0, 0, \frac{\partial P_2'}{\partial \lambda_1}, \frac{\partial P_2'}{\partial \lambda_2}) ,$$

$$d(\lambda_1^2+\lambda_2^2-1) = (\ldots, 0, 0, 2\lambda_1, 2\lambda_2) ,$$

and, since

$$- \det \begin{pmatrix} \lambda_1 & 1 & 0 & 0 \\ 1 & 0 & \frac{\partial P_1'}{\partial \lambda_1} & \frac{\partial P_1'}{\partial \lambda_2} \\ 0 & 0 & \frac{\partial P_2'}{\partial \lambda_1} & \frac{\partial P_2'}{\partial \lambda_2} \\ 0 & 0 & 2\lambda_1 & 2\lambda_2 \end{pmatrix} = 2 \left[\lambda_2 \frac{\partial P_2'}{\partial \lambda_1} - \lambda_1 \frac{\partial P_2'}{\partial \lambda_2} \right] =$$

$$= 2 \left[\lambda_2 \frac{\partial P_2'}{\partial \lambda_1} + \lambda_1 \frac{\partial P_1'}{\partial \lambda_1} \right] = 2 \operatorname{Re} (\lambda^{-1} P''(\lambda)) .$$

Thus, it suffices to prove that for a dense subset of V , $\operatorname{Re} (\lambda^{-1} P''(\lambda)) \neq 0$.

It is obvious that the set of those $x \in V$ for which $P''(\lambda) \neq 0$ is dense in V . If λ is real and $\lambda \in S$,

$P''(\lambda) \neq 0$, then also $\lambda^{-1}P''(\lambda) = \mathrm{Re}\,\lambda^{-1}P''(\lambda) \neq 0$.

Assume that λ is not real, $\lambda \in S$ and $P''(\lambda) \neq 0$. Then $\lambda^{-1}P''(\lambda) = \bar{\lambda}P''(\lambda) = \bar{\lambda}(\lambda - \bar{\lambda})^2 R(\lambda)$, where $R(\mu)$ is real for μ real. For ε real denote $P_\varepsilon(\mu) = (\mu-\lambda)^2(\mu-\bar{\lambda})^2[R(\mu)+\varepsilon] = \mu^n + \alpha_{1\varepsilon}\mu^{n-1} + \ldots + \alpha_{n\varepsilon}$. $P_\varepsilon(\mu)$ is real for μ real and $(\alpha_{1\varepsilon}, \ldots, \alpha_{n\varepsilon}, \lambda_1, \lambda_2) \in V$. We have $\mathrm{Re}(\bar{\lambda}P_\varepsilon''(\lambda)) - \mathrm{Re}(\bar{\lambda}P''(\lambda)) = \varepsilon\,\mathrm{Re}[\bar{\lambda}(\lambda-\bar{\lambda})^2] = -4\varepsilon\lambda_1\lambda_2$. Since both $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, there is an $\varepsilon > 0$ arbitrarily small such that $\mathrm{Re}[\bar{\lambda}P_\varepsilon''(\lambda)] \neq 0$. This proves the density in V of the set of points x for which $\mathrm{Re}(\lambda^{-1}P''(\lambda)) \neq 0$.

Let i be such that $\tilde{\pi}(\mathcal{M}_1) \cap V_i \neq \emptyset$, $\tilde{\pi}(\mathcal{M}_1) \cap V_j = \emptyset$ for $j < i$. Since $\bigcup_{j=1}^{i} V_j$ is open, $\mathcal{M} = \tilde{\pi}^{-1}(V_i) = \tilde{\pi}^{-1}(\bigcup_{j=1}^{i} V_j)$ is open in $\mathcal{M}_1$ and, by Lemma 3, $\pi(\mathcal{M}_0)$ is open in V_i . From this and the Sard's theorem ([6, Theorem 15.1]) it follows that there is a point $\tilde{A} \in \mathcal{M}_0$ at which $\tilde{\pi}$ is regular. Thus, locally $\tilde{\pi}^{-1}(\tilde{\pi}(\tilde{A}))$ is an imbedded submanifold of the dimension $\dim \mathcal{M}_1 - \dim V_i \geq \dim \mathcal{M}_1 - n + 2$. On the other hand, from Corollary 2 it follows $\dim \tilde{\pi}^{-1}(\tilde{\pi}(\tilde{A})) \leq n^2 - n$. Consequently, $\dim \mathcal{M}_1 \leq n^2 - 2$, q.e.d.

Lemma 4. If $\lambda_{20} \neq 0$, then $\mathrm{codim}\, \mathcal{N}_1 \geq 4$.

The proof of this lemma is similar to that of Lemma 1, with V replaced by the set $W \subset R^{n+2}$ of points $(\alpha_1, \ldots, \alpha_n, \lambda_{10}, \lambda_{20})$ for which $\lambda_0 = \lambda_{10} + i\lambda_{20}$ is a root of $P(\lambda) = \lambda^n + \alpha_1\lambda^{n-1} + \ldots + \alpha_n$.

This is again an algebraic variety defined by the equations $\lambda_1 - \lambda_{10} = \lambda_2 - \lambda_{20} = 0, P_1(\lambda_1, \lambda_2) = P_2(\lambda_1, \lambda_2) = 0$. The differentials of the polynomials at the points of W are

$$dP_1 = (\ldots, \lambda_{10}, 1, \frac{\partial P_1}{\partial \lambda_1}, \frac{\partial P_1}{\partial \lambda_2}),$$

$$dP_2 = (\ldots, \lambda_{20}, 0, \frac{\partial P_2}{\partial \lambda_1}, \frac{\partial P_2}{\partial \lambda_2}),$$

$$d(\lambda_1 - \lambda_{10}) = (\ldots, 0, 0, 1, 0),$$

$$d(\lambda_2 - \lambda_{20}) = (\ldots, 0, 0, 0, 1).$$

Obviously, they are independent if $\lambda_{20} \neq 0$. The rest of the proof is analogous to the proof of Lemma 1.

Proof of Proposition 1. Openness follows from the fact that both $\mathcal{U}_1$ and $\mathcal{U}_2$ are closed.

For the proof of density we consider the sets $\widetilde{\mathcal{U}}_1, \widetilde{\mathcal{U}}_2(\lambda_{10}, \lambda_{20})$ with $\lambda_{20} \neq 0$ and the space $\widetilde{\Phi}$ of maps $F : int\, I \times R^2 \longrightarrow \widetilde{\mathcal{U}}$, defined by $\widetilde{F} = F|_{int\, I} \times id$, $F \in \Phi$, endowed with the C^r uniform topology. Further, we denote by $\widetilde{\Psi}_i = \{\widetilde{F} \mid \widetilde{F}(I) \cap \cap \bigcup_{j=r-i+1}^{r} M_j = \emptyset\}$ for $1 \leq i \leq r$, $\widetilde{\Psi}_{r+i} = \{\widetilde{F} \mid \widetilde{F}(I) \cap \cap \widetilde{\mathcal{U}}_1 \cap \bigcup_{j=s-i+1}^{s} N_j = \emptyset\}$ for $1 \leq i \leq s$. Since Ψ_ℓ is the intersection of the projections of $\widetilde{\Psi}_{r+s}$ taken over all nonreal ℓ-th roots of unity, it suffices to prove that $\widetilde{\Psi}_{r+s}$ is dense in $\widetilde{\Phi}$. We prove this by induction showing that every $\widetilde{F} \in \widetilde{\Psi}_i$ can be approximated arbitrarily closely by an $\widetilde{F}' \in \widetilde{\Psi}_{i+1}$. Without loss

of generality we assume $1 < i < n$.

The map $\varphi : \Phi \longrightarrow \widetilde{\Phi}$ given by $\varphi(F) = \widetilde{F}$ is a C^n-representation (here and further in this proof we use the terminology of [6]) and the evaluation map meets M_{n-i} transversally. Due to the dimension estimates of Lemma 1 and Lemma 4, the existence of the approximation of F not intersecting M_{n-i} follows from the transversality theorem [6, Theorem 19.1] and the openness of $\widetilde{\Psi}_i$, q.e.d.

Denote $\mathcal{U}_3$ the subset of $\mathcal{U}$ consisting of matrices having an eigenvalue on S . Again, we associate with $\mathcal{U}_3$ the algebraic variety $\widetilde{\mathcal{U}}_3$ in $\widetilde{\mathcal{U}}$, defined by

$$\widetilde{\mathcal{U}}_3 = \{(A, \lambda_1, \lambda_2) \mid P_1(\lambda_1, \lambda_2) = P_2(\lambda_1, \lambda_2) = \lambda_1^2 + \lambda_2^2 - 1 = 0\}$$

whose projection is $\mathcal{U}_3$. Thus, $\widetilde{\mathcal{U}}_3 = \bigcup_{i=1}^{n} \mathcal{K}_i$, where $\mathcal{K}_i$ are mutually disjoint manifolds of decreasing dimension and $\bigcup_{j=i}^{n} \mathcal{K}_j$ is closed in $\widetilde{\mathcal{U}}_3$ for every i .

Lemma 5. $\operatorname{codim} \mathcal{K}_1 = 3$.

Proof. The proof of the inequality $\dim \mathcal{K}_1 \geq 3$ is analogous to that of Lemma 1. We only note that the differentials of the defining polynomials $P_1, P_2, \lambda_1^2 + \lambda_2^2 - 1$ of $\tilde{\pi}(\widetilde{\mathcal{U}}_3) \subset \mathbb{R}^{n+2}$ ($\tilde{\pi}$ defined as in Corollary 2) are independent if $\operatorname{Re}(\lambda P'(\lambda)) \neq 0$; it can be shown similarly as in the proof of Lemma 1 that this is true for a dense subset of $\tilde{\pi}(\widetilde{\mathcal{U}}_3)$.

To prove the opposite inequality assume $I = [0, 2]$ and consider the map $F(t) = \operatorname{diag}\{t, 0, \ldots, 0\}$. If

$\operatorname{codim} \mathcal{K}_1 < 3$ then it would follow from the transversality argument used in the proof of Proposition 1 that there should exist a small C^r perturbation $\hat{F}$ of F no value of which would have an eigenvalue on S. This, however, is obviously impossible.

Proposition 2. Let $J \subset I$ be a closed interval, $J \subset \operatorname{int} I$. Then, for every $l > 2$ the subset $\Psi_l^0(J) \subset \Psi_l(J)$ of all F such that F meets $\tilde{\mathcal{U}}_3$ transversally (i.e. F meets transversally $\mathcal{K}_1$ and does not meet $\mathcal{K}_i$ for $i > 1$ at all) is open dense in $\Psi_l(J)$, and, thus, in Φ.

The proof is analogous to that of Proposition 1.

Corollary 3. Given J as in Proposition 2, the set $\Psi^0(J)$ of maps $F \in \Phi$ such that $F(J) \cap (\mathcal{U}_1 \cup \mathcal{U}_2) = 0$ and F meets $\tilde{\mathcal{U}}_3$ transversally over J is residual in Φ.

Lemma 6. Let $F \in \Phi$ and let λ_0 be a simple eigenvalue of $F(t_0)$, where $t_0 \in I$. Then there is a neighbourhood N of t_0 in I and a unique function $\lambda : N \to C$ such that $\lambda(t_0) = \lambda_0$ and $\lambda(t)$ is an eigenvalue of $F(t)$ for $t \in N$. Further, there is a nonsingular C^r matrix $C(t)$ on N such that $C^{-1} F C = B$, where the first column of $B(t)$ is the transpose of $(\lambda(t), 0, \ldots, 0)$.

Proof. Without loss of generality we may assume that $F(t_0)$ is in the Jordan canonical form with λ_0 in the first column. Choose $C(t_0) = E$ (the unity matrix) and $C(t) = (c_1(t), \ldots, c_m(t))$, $\lambda(t)$ as the solution of

the set of equations $F(t)c_1(t) = \lambda(t)c_1(t)$, $c_i(t) = c_i(t_0), i > 1, |c_1(t)| = 1$ ($|.|$ being the Euclidean norm). It is easy to check that the Jacobian of this set of equations at t_0 is not zero. The implicit function theorem completes the proof.

Remar 1. Under the assumptions of Lemma 6, for λ_0 not real, starting from the real canonical form of $F(t_0)$, one can similarly prove that there is a C^r real matrix $C(t)$ in some neighbourhood of t_0 in I that brings $F(t)$ into the form

$$\begin{pmatrix} B_1(t), B_2(t) \\ 0 \quad , B_3(t) \end{pmatrix}, \text{ where } B_1(t) = \begin{pmatrix} \operatorname{Re}\lambda(t), \operatorname{Im}\lambda(t) \\ -\operatorname{Im}\lambda(t), \operatorname{Re}\lambda(t) \end{pmatrix}.$$

Corollary 4. Let $F \in \Phi$, $t_0 \in I$ and let $\lambda_{10}, \ldots, \lambda_{k0}$ be simple eigenvalues of $F(t_0)$. Then, there is a neighbourhood N of t_0 in I and unique C^r functions $\lambda_i : N \longrightarrow C$ such that $\lambda_i(t_0) = \lambda_{i0}$ and $\lambda_i(t)$ are eigenvalues of $F(t)$ for $t \in N$. Further, there is a C^r matrix $C(t)$ on N such that $C^{-1}AC = B$, where B has the form $\begin{pmatrix} B_1 , B_2 \\ 0 , B_3 \end{pmatrix}$ and B_1 is triangular with $\lambda_1, \ldots, \lambda_k$ on the diagonal. Also, there is a real C^r matrix $\hat{C}(t)$ on N that brings $F(t)$ into the form $\begin{pmatrix} \hat{B}_1(t), \hat{B}_2(t) \\ 0 \quad , \hat{B}_3(t) \end{pmatrix}$, where $\hat{B}_1(t)$ is block diagonal with blocks as in Remark 1.

Proposition 3. Let $F \in \mathcal{Y}^0_\ell(J)$ for some $\ell > 2$. Then, the eigenvalues of F meet S transversally.

By this proposition we mean that the functions λ, defined in Lemma 6 for $\lambda_0 \in S$ (note that such λ_0 are simple) meet S transversally.

Proof. Let $\lambda(t_0) \in S$ be an eigenvalue of $F(t_0)$. By Lemma 6, there is a nonsingular C^n matrix $C(t)$ such that $C^{-1}(t)F(t)C(t) = B(t)$, where $B(t)$ has the form specified in Lemma 6. Denote $B(t, \mu)$ the matrix obtained from $B(t)$ by replacing in the first column $\lambda(t)$ by μ. Denote by $\mu(t)$ the orthogonal projection of $\lambda(t)$ on S, φ the Euclidean distance. Since $C(t)B(t, \mu(t))C^{-1}(t) \in \mathcal{U}_3$ and $\mathcal{K}_1$ is open in $\widetilde{\mathcal{U}}_3$, $(C(t)B(t, \mu(t))C^{-1}(t), \mu_1(t), \mu_2(t)) \in \mathcal{K}_1$, for t sufficiently close to t_0, where $\mu = \mu_1 + i\mu_2$. We have

$$|\lambda(t)| - 1 = |\lambda(t) - \mu(t)| = \varphi(B(t), B(t, \mu(t))) \geq |C(t)|^{-1},$$

$$|C(t)^{-1}|^{-1}\varphi(F(t), C(t)B(t, \mu(t))C^{-1}(t)) \geq k_1 \varphi(\widetilde{F}(t), \mathcal{K}_1),$$

where $k_1 > 0$ is a suitable constant. If $\widetilde{F}$ meets $\mathcal{K}_1$ transversally, then obviously $\varphi(\widetilde{F}(t), \mathcal{K}_1) \geq k_2 |t - t_0|$ for some $k_2 > 0$. Consequently, $\left.\frac{d|\lambda(t)|}{dt}\right|_{t=t_0} \neq 0$, q.e.d.

Corollary 5. The number of such $t \in J$ for which an eigenvalue of $F(t)$ is on S, is finite for every $F \in \mathcal{F}_\ell^0(J)$.

Theorem 1. Let $J \subset \mathrm{int}\, I$ be a closed interval. Then, the set $\Phi_{1\ell}(J)$ of those $F \in \Phi$, satisfying

(i) $F(t)$ has no double eigenvalue on S,

(ii) $F(t)$ has no non-real ℓ-th root of unity as ei-

genvalue,
(iii) the eigenvalues of $F(t)$ meet S transversally,
(iv) if an eigenvalue of $F(t)$ lies on S, then no other eigenvalue of $F(t)$ lies on S except of its complex conjugate,
for every $t \in J$, is open dense in Φ.

Corollary 6. The set $\Phi_1(J)$ of those $F \in \Phi$ satisfying (i),(iii),(iv) of Theorem 1 and such that for every $t \in J$, $F(t)$ has no non-real root of unity as eigenvalue, is residual in Φ.

Proof. Openness is obvious. From Propositions 1 - 3 it follows that the set of maps from Φ, satisfying (i) - (iii) (i.e. the set $\Psi_\ell^0(J)$), is open dense in Φ. Therefore, it suffices to prove that every $F \in \Psi_\ell^0(J)$ can be arbitrarily closely approximated by an $\hat{F} \in \Psi_\ell^0(J)$ satisfying (iv). In virtue of Corollary 4 it suffices to show that if for some t_0 (iv) is not satisfied it is possible to perturb F in an arbitrary small neighbourhood N of t_0 by an arbitrary small perturbation, without changing it outside N, in such a way that (i) - (iv) will be true for the perturbation of F for every $t \in N$.

Assume that for some $t_0 \in J$, k pairs of conjugate eigenvalues λ_j^0, $\overline{\lambda_j^0}$, $j = 1, \dots, k$ lie on S (the modification of the proof for the case of some eigenvalue being real is straightforward). Let α be so small that the functions λ_j, defined by λ_j^0, t_0 as in Lemma 6 exist and do not meet S except at t_0 and no other eigenvalue of $F(t)$ lies on S on $K \cap J$, where

$K = [t_0 - \alpha, t_0 + \alpha]$, and that there is a C^n matrix C such that $C^{-1}(t)F(t)C(t) = B(t)$ has the form

$$B = diag\left\{\begin{pmatrix} \lambda_{11}, & \lambda_{12} \\ -\lambda_{21}, & \lambda_{22} \end{pmatrix}, \dots, \begin{pmatrix} \lambda_{k1}, & \lambda_{k2} \\ -\lambda_{k2}, & \lambda_{k1} \end{pmatrix}, B_1\right\}$$

where $\lambda_j = \lambda_{j1} + i\lambda_{j2}$ (cf. Remark 1). Choose an $\varepsilon < \frac{\alpha}{2}$, k real mutually distinct numbers τ_j, $j = 1, \dots, k$ such that $|\tau_j| < \varepsilon$ and a bump function $\chi : N \to R$ such that $\chi(t) = 0$ outside K, $\chi(t) = 1$ for $t \in K_0 = [t_0 - \frac{\alpha}{2}, t_0 + \frac{\alpha}{2}]$, $\hat{\lambda}_j(t) = \lambda_j(t + \tau_j \chi(t))$,

$$\hat{B}(t) = diag\left\{\begin{pmatrix} \hat{\lambda}_{11}(t) & \hat{\lambda}_{12}(t) \\ -\hat{\lambda}_{21}(t) & \hat{\lambda}_{11}(t) \end{pmatrix}, \dots, \begin{pmatrix} \hat{\lambda}_{k1}(t), & \hat{\lambda}_{k2}(t) \\ -\hat{\lambda}_{k2}(t), & \hat{\lambda}_{k1}(t) \end{pmatrix}, B_1(t)\right\},$$

$$F(t) = \begin{cases} F(t) & \text{for } t \notin K, \\ C(t)\hat{B}(t)C^{-1}(t) & \text{for } t \in K. \end{cases}$$

It is obvious that $\hat{F} \in \Psi_\ell^0$ and, in $K \cap J$, $\hat{\lambda}_j$ meets S exclusively at the point $t_0 - \tau_j$. If τ_j are chosen small enough, $\hat{F}$ will be arbitrarily close to F, q.e.d.

§ 3

In [1, § 2] it was shown that for $f \in \mathcal{F}_1$, each point of $\overline{Z} \setminus Z_k$ (such points have been called branching points) is contained in some set Z_ℓ with ℓ being a di-

visor of k and that some eigenvalue of df_p^{ℓ} at such point has to be a root of unity different from 1 .

Theorem 2. There is a subset $\mathcal{F}_2$ of $\mathcal{F}_1$, residual in $\mathcal{F}$ such that for every $f \in \mathcal{F}_2$, the following is true for every $(p_0, m_0) \in Z_k(f)$, $k \geq 1$:

(i) $df_{p_0}{}^{k}(m_0)$ has no double eigenvalue on S ,

(ii) $df_{p_0}{}^{k}(m_0)$ has no non-real root of 1 as an eigenvalue .

(iii) The eigenvalues of $df_p{}^{k}(m)$ meet S transversally at (p_0, m_0) .

(iv) If an eigenvalue of $df_{p_0}{}^{k}(m_0)$ lies on S , then there is no other eigenvalue of $df_{p_0}{}^{k}(m_0)$ on S except of its complex conjugate.

Corollary 7. For $f \in \mathcal{F}_2$, $(p, m) \in Z_k(f)$ can be a branching point only if one of the eigenvalues of $df_p(m)$ is -1 , the other being outside S .

Remark 2. Denote $\mathcal{F}_{2k\ell}$ the subset of $\mathcal{F}_{1k}$ of those mappings, satisfying (i),(iii),(iv) for $1 \leq k \leq h$ and (ii) with "roots" replaced by " ℓ -th roots" for $1 \leq k \leq h$. Then, $\mathcal{F}_{2k\ell}$ is open dense in $\mathcal{F}$.

Remark 3. (iii) should be understood as follows: If an eigenvalue λ_0 of $df_{p_0}{}^{k}(m_0)$ is on S , then in some neighbourhood N of (p_0, m_0) in Z_k , there is a unique C^n function $\lambda : N \longrightarrow C$ such that $\lambda(p, m)$ is

an eigenvalue of $df_p^k(m)$ for $(p,m) \in N$ and $\lambda(p_0, m_0) = \lambda_0$. This λ meets S transversally.

Proof. It suffices to prove Remark 2, from which Theorem 2 follows. We carry out the proof for $k = 1$, i.e. we prove that $\mathcal{F}_{21\ell}$ is open dense for any ℓ; the extension for $k > 1$ is similar as in the proof of [1, Theorem 1].

The openness of $\mathcal{F}_{21}$ is obvious. To prove density, assume $f \in \mathcal{F}_{11}$. Then, by [1, Theorem 1], there is an open set U containing $X_1(f)$ such that for every $(p_0, m_0) \in U$, (i) - (iv) is trivially satisfied. $Z_1 \setminus U$ can be covered locally finitely by a countable family $(W_\alpha, \mu_\alpha \times x_\alpha)$, $W_\alpha = U_\alpha \times V_\alpha$ of coordinate neighbourhoods in such a way that for any $K \in P \times M$ compact, $W_\alpha \cap K \neq \emptyset$ for a finite number of α's only and $(W_\alpha, \mu_\alpha \times x_\alpha)$ satisfy (iv) of [1, Theorem 1] (i.e. $W_\alpha \cap Z_1$ is the graph of a C^k function $\varphi_\alpha : U \to V$). We show how for any open W'_α, $\overline{W}'_\alpha \subset \overline{W}'_\alpha = U'_\alpha \times V'_\alpha$, f can be approximated by $\hat{f}$ such that $\hat{f}$ coincides with f outside W_α and satisfies (i) - (iv) of Theorem 2 for every $(p_0, m) \in Z_1 \cap W_\alpha$. The construction of an approximation of f satisfying (i) - (iv) for any $(p_0, m_0) \in Z_1$ is then standard. In the rest of the proof we drop the subscript α.

In the coordinates $(p, m) \mapsto (\mu, y)$, $y = x - x_0 \varphi(p)$, f can be represented by

$$y' = A(\mu) y + Y(\mu, y)$$

where the primed coordinates are those of the image,

$Y(\mu, 0) = 0$, $dY(\mu, 0) = 0$.

By Theorem 1, we can approximate $A : \mu(U) \longrightarrow \mathcal{U}$ by a map $\hat{A} : \mu(U) \longrightarrow \mathcal{U}$ such that A satisfies (i) - (iv) of Theorem 1 on U.

Let $\psi : (\mu \times x)(W) \longrightarrow R$ be a C^r bump function such that $\psi = 1$ on $(\mu \times x)(\overline{W'})$ and $\psi = 0$ outside $(\mu \times x)(W)$. Denote by $\hat{f}$ the map which coincides with f outside W and is given in W by the coordinate representation

$$y' = [A(\mu) + \psi(\mu, y)(\hat{A}(\mu) - A(\mu))]y + Y(\mu, y).$$

If we choose A sufficiently close to A, $\hat{f}$ will be arbitrarily close to f and will satisfy (i) - (iv) for every $(n_0, m_0) \in W'$.

Denote by Y_k the set of points $(n, m) \in Z_k$ for which one eigenvalue of $df_n^k(m)$ is -1. For $(n, m) \in Z_k$ denote $h(n, m)$ the number of eigenvalues of $df_n^k(m)$ with modulus less than 1.

Theorem 3. Assume $r > 2$. Then, there is a subset $\mathcal{F}_3$ of $\mathcal{F}_2$, residual in $\mathcal{F}$, such that every $f \in \mathcal{F}_3$ has the following properties:

(i) Y_k coincides with the set of k-periodic branching points,

(ii) for every $(n_0, m_0) \in Y_k$, there is a coordinate neighbourhood $(W, \mu \times x)$, $W = U \times V$ of (n_0, m_0) such that $\mu(n_0) = 0$, $x(m_0) = 0$, $Z_k \cap W = U \times \{0\}$ and

(a) $Z_{2k} \cap W$ consists of two components, separa-

ted by (p_0, m_0); all points $(p, m) \in Z_{2k} \cap W$ satisfy $\mu(p) > 0$ and $Z_{2k} \cap W \cup \{(p_0, m_0)\}$ is a C^1 (but not C^2) submanifold of W.

(b) No eigenvalue of $[(Z_k \cup Z_{2k}) \cap W] \setminus \{(p_0, m_0)\}$ is on S; either $h(p, m) = h(p', m') = h(p'', m'') + 1$ or $h(p, m) = h(p', m') = h(p'', m'') - 1$ for any $(p, m) \in Z_k \cap W$, $\mu(p) < 0$, $(p', m') \in Z_{2k} \cap W$, $(p'', m'') \in Z_k \cap W$, $\mu(p'') > 0$,

(c) $W \setminus (Z_k \cup Z_{2k})$ contains no invariant set.

Proof. Again, we carry out the proof for $k = 1$, the proof of its extension for $k > 1$ being as in [1, Theorem 1].

Let $f \in \mathcal{F}_{21\ell}$. Then, $Y_1(f)$ is discrete and, if $(p_0, m_0) \in Y_1$, one eigenvalue of $df_{p_0}(m_0)$ is -1 and the remaining ones can be divided into two groups according to whether their moduli are < 1 or > 1, the number of the former ones being $h(p_0, m_0)$. Thus, using [6, Appendix 3] as in [1, Lemma 4], it follows that we can choose the coordinates (μ, x) in such a way that $x = (x_1, y, z)$, $\dim x_1 = 1$, $\dim y = h(p_0, m_0)$ and the coordinate representation of f in these coordinates is as follows:

$$
\begin{aligned}
x_1 &= -x_1 + \alpha\mu x_1 + \beta x_1^2 + \gamma x_1^3 + \omega(\mu, x_1, y, z), \\
y &= Ay + Y(\mu, x_1, y, z), \\
z &= Cz + Z(\mu, x_1, y, z),
\end{aligned} \tag{3}
$$

where ω, Y, Z are C^∞ and

ω, Y, Z are C^r and $Y(\mu, x_1, 0, x) = 0, Z(\mu, x_1, y, 0) = 0$,
$\omega(\mu, x_1, y, x) = 0(|x_1^3| + |\mu x_1| + |y| + |x|)$,
$d\omega(0,0,0,0) = 0$,
$dY(0,0,0,0) = 0, dZ(0,0,0,0) = 0$.

We denote by $\mathcal{F}_{31}$ the subset of $\mathcal{F}_{11}$ of those maps in the coordinate representation (3) of which $\beta^2 + \gamma \neq 0$ for every $(\mu_0, m_0) \in Y_1(f)$. The definition of $\mathcal{F}_{31}$ does not depend on the choice of particular coordinates and the set $\mathcal{F}_{31}$ is open dense in $\mathcal{F}$. The proof of this as well as the proof that the maps of $\mathcal{F}_{31}$ satisfy (i),(ii) for $k = 1$ does not differ from the corresponding part of the proof of [1, Theorem 3], except of the proof of (ii)(c), where, because of the possible presence of the eigenvalues of moduli both < 1 and > 1 one has to use the argumentation of the proof of [1, Lemma 4].

As a corollary of [1, Theorem 1] and Theorem 3 we obtain

Theorem 4. Assume $r > 2$. Then, for every $f \in \mathcal{F}_3$:
(i) for k odd, Z_k is a closed submanifold of $P \times M$,
(ii) for k even, either Z_k is closed and $Y_{k/2}$ is empty, or Z_k is a C^1 (but not C^2) submanifold of $P \times M$ and $\overline{Z}_k \setminus Z_k$ is discrete and coincides with $Y_{k/2}$.

Remark 4. This theorem corrects the erroneous formulation of its two dimensional version [1, Theorem 4], in which the possibility of Z_k being closed was omitted.

R e f e r e n c e s :

[1] P. BRUNOVSKÝ: On one-parameter families of diffeomorph-

isms, Comment.Math.Univ.Carolinae 11(1970), 559-581.

[2] M.M. PEIXOTO: On an approximation theorem of Kupka and Smale, Journal of Differential Equations 3 (1966), 214-227.

[3] H. WHITNEY: Elementary structure of real algebraic varieties, Annals of Mathematics 66(1957), 545-556.

[4] R. THOM, H. LEVINE: Singularities of differentiable mappings, Russian translation, Mir,Moscow, 1969.

[5] F.R. GANTMACHER: Teorija matric, Nauka,Moscow, 1966.

[6] R. ABRAHAM, J. ROBBIN: Transversal mappings and flows, Benjamin, 1967.

Matematický ústav SAV
Bratislava
Československo

(Oblatum 28.4. 1971)

P. Brunovský, B. Fiedler

Numbers of zeros on invariant manifolds in reaction-diffusion equations

Nonlinear Anal. 10(2) (1986), 179–193.

Nonlinear Analysis, Theory, Methods & Applications, Vol. 10, No. 2, pp. 179–193, 1986.
Printed in Great Britain.

0362–546X/86 $3.00 + .00

NUMBERS OF ZEROS ON INVARIANT MANIFOLDS IN REACTION–DIFFUSION EQUATIONS

PAVOL BRUNOVSKÝ
Univerzita Komenského, Institute of Applied Mathematics, Mlynská dolina, 84215 Bratislava 2, ČSSR

and

BERNOLD FIEDLER
Sonderforschungsbereich 123, Universität Heidelberg, Im Neuenheimer Feld 293, 6900 Heidelberg, FRG

(*Received* 10 *November* 1984; *received for publication* 9 *April* 1985)

Key words and phrases: Lap number, zero number, invariant manifolds, reaction–diffusion equation.

INTRODUCTION

CONSIDER the one-dimensional reaction–diffusion equation

$$u_t = u_{xx} + f(x, u), t > 0, 0 < x < 1 \tag{0.1}$$

with the Dirichlet boundary conditions

$$u(t, 0) = u(t, 1) = 0, \tag{0.2}$$

where f: $[0, 1] \times \mathbb{R} \to \mathbb{R}$ is $BC^1 \cap C^\kappa$, $\kappa > 1$. The equations (0.1), (0.2) can be viewed as a particular case of the abstract equation

$$du/dt + Au = f(u) \tag{0.3}$$

in a Banach space X, the basic theory of which is developed in [5]. For (0.1), (0.2), $X = L^2[0, 1]$, A is the closure of the operator defined by $Av = -v''$ for $v \in C^2[0, 1]$, $v(0) = v(1) = 0$, F: $L^2[0, 1] \to L^2[0, 1]$ is given by $F(v)(x) = f(x, v(x))$. We frequently work in the Hilbert space $X^1 = \mathcal{D}(A) = H_0^1([0, 1]) \cap H^2([0, 1])$ with F: $X^1 \to X^1$ also being C_κ. Let $|\cdot|$ denote the norm on X^1.

Applying the results of [5] one obtains that (0.1), (0.2) generates a local semiflow S on X^1. The semiflow S is a continuous map of an open neighbourhood U of $\{0\} \times X^1$ in $\mathbb{R}^+ \times X^1$ into X^1 defined by

$$S_t(v)(x) = u(t, x) \quad \text{for} \quad (t, v) \in U,$$

where u is the solution of (0.1), (0.2), satisfying

$$u(0, x) = v(x) \quad \text{for} \quad 0 < x < 1. \tag{0.4}$$

It has the properties $S_0(v) = v$, $S_{t+s}(v) = S_t \circ S_s(v)$ as long as (s, v) and $(t, S_s(v))$ are in U [5]. In order not to obscure the formulations by technicalities we shall assume that S is a global semiflow, i.e. $U = \mathbb{R}^+ \times X^1$. This is by no means an essential restriction; sufficient conditions can be found in [5, Chapter 3].

The critical points of S are the stationary solutions of (0.1), (0.2), i.e. the solutions of the equation

$$v'' + f(x, v) = 0, v(0) = v(1) = 0. \tag{0.5}$$

The qualitative properties of S near a critical point v are determined by the linearization of (0.1), (0.2) at v which is the equation

$$y_t = y_{xx} + f_u(x, v(x))y \tag{0.6}$$

$$y(t, 0) = y(t, 1) = 0. \tag{0.7}$$

The solution v is called hyperbolic if 0 is not an eigenvalue of the operator $L = A - F'(v)$, i.e. (0.6), (0.7) do not admit a nontrivial stationary solution y.

An important information about the global structure of the semiflow of (0.1), (0.2) is given by the orbit connections of different stationary solutions [2, 3, 5]. By a connecting orbit of the stationary solutions v_1, v_2 we understand a solution u of (0.1), (0.2) which exists for all $t \in (-\infty, \infty)$ and satisfies

$$\lim_{t \to -\infty} u(t, x) = v_1(x), \lim_{t \to \infty} u(t, x) = v_2(x)$$

in $H^2[0, 1]$. In the terminology of [5], $u(t, \cdot)$ has to be in the stable manifold of v_2 and the unstable manifold of v_1, provided v_1, v_2 are hyperbolic.

In this paper we obtain estimates on the number of zeros (or, more precisely, the zero number defined below) of $u(t, \cdot)$-v_1 and $u(t, \cdot)$-v_2. This information can be used to conclude existence and nonexistence of connections. Our approach provides an alternative to the (slightly different) zero number of u_t, which was used by Hale and Nascimento [3] to solve the connection problem for f of the Chafee–Infante type (see e.g. [5, Section 5.3]).

For any continuous function ϕ: $[0, 1] \to \mathbb{R}$ we define the *zero number* $z(\phi)$ as follows. Let $n \geqq 0$ be the maximal element of $\mathbb{N}_0 \cup \{\infty\}$ such that there is a strictly increasing sequence $0 \leqq x_0 < x_1 < \ldots < x_n \leqq 1$ with $\phi(x_j)$ of alternating signs:

$$\phi(x_j) \cdot \phi(x_{j+1}) < 0 \quad \text{for } 0 \leqq j < n.$$

If n is finite let $z(\phi) := n$, and $z(\phi) := \infty$ otherwise. Note that we put $z(0) := 0$.

As a first example consider the linearized equation (0.6), (0.7). The operator $L = A - F'(v)$ has eigenvalues $\lambda_0 < \lambda_1 < \ldots$ with eigenfunctions $\phi_0, \phi_1, \ldots$. By Sturm–Liouville theory $z(\phi_k) = k$ and indeed it is a classical result (see [0, p. 549]) that for $0 \leqq i < j < \infty$

$$i \leqq z(\phi) \leqq j, \tag{0.8}$$

whenever ϕ is a (nontrivial) linear combination of $\phi_i, \ldots, \phi_j$. As a trivial illustration of our approach we prove estimate (0.8) in corollary 1.2, using the dynamic equation (0.6), (0.7).

All our results depend on a basic observation, lemma 1.1, going back to Redheffer, Walter [8] and, more recently, Matano [6]. According to lemma 1.1,

$$z(u(t, \cdot)) \text{ is nonincreasing}$$

as a function of time t along solutions of equation (0.1), (0.2) provided that f satisfies the condition.

$$f(x, 0) = 0 \quad \text{for } 0 < x < 1. \tag{0.9}$$

The proof is elementary and relies on the maximum principle for parabolic equations. For the convenience of the reader we present it in detail below.

In the nonlinear case, let v be a hyperbolic stationary solution of (0.1), (0.2). Then the eigenvalues λ_j of the linearized equation with corresponding eigenfunctions ϕ_j satisfy $\lambda_0 < \ldots < \lambda_{n-1} < 0 < \lambda_n < \ldots$ for some $n \geqq 0$. Further by [5, theorems 5.2.1, 6.1.9] there exist immersed invariant C^K-manifolds W^u and $W^s \subset X^1$ of the flow S through $v = 0$ with the properties:

(i) for $w \in W^u$ (resp. W^s) the solution $u(t, \cdot) = S(t)w$ exists for all real t and satisfies $\lim u(t, \cdot) = v$ as $t \to -\infty$, (resp. $t \to +\infty$);

(ii) the tangent space of W^u (resp. W^s) at v is spanned by the ϕ_k with $k < n$ (resp. $k \geqq n$).

W^u is called the unstable manifold and W^s the stable manifold of v.

Our main result, given in Sections 2 and 3, states that

$$z(w - v) < \dim W^u \quad \text{for} \quad w \in W^u \tag{0.10}$$

(theorem 2.1) and

$$z(w - v) \geqq \dim W^u \quad \text{for} \quad w \in W^s \backslash \{v\} \tag{0.11}$$

(theorem 3.2). Note that these estimates are suggested by the respective tangent spaces of W^u and W^s, together with the Sturm–Liouville estimate (0.8).

The crucial observation of our proof is that for $v \equiv 0$:

$$\lim \frac{u(t)}{|u(t)|} = \phi_k \tag{0.12}$$

—for $t \to -\infty$ on W^u and some $k < n$

—for $t \to +\infty$ on W^s and some $k \geqq n$, provided that $z(u(t, \cdot))$ is eventually finite.

Actually it is quite simple to prove (0.12) on W^u, as we will indicate at the end of Section 2. However, analysis on the infinite dimensional stable manifold W^s is quite delicate and we need detailed information on the fine structure of W^s before we can prove (0.12). For illustration we pursue an analogous approach to W^u in Section 1, as a preparation to the stable manifold case.

1. COUNTING ZEROS

In the introduction we defined the zero number $z(\phi)$ of a continuous real function ϕ as the maximal number of sign changes of ϕ. In this section we show that z decreases along solutions $u(t, \cdot)$ of the parabolic equation (0.1) with Dirichlet boundary conditions, assuming that

$$f(x, 0) = 0 \quad \text{for all} \quad x \in I, \tag{1.1}$$

$I := [0, 1]$. This result is essentially in [8, corollary 3] who consider $f = f(t, x, u_x, u_{xx})$ independent of u. Similarly, Matano [6] investigates the lap number of ϕ, which is the zero number of ϕ_x and was called "maximum order of a saw in ϕ" by Redheffer and Walter [8].

Note that by definition the function

$$z\colon C^0(I) \to \mathbb{N} \cup \{\infty\}$$

is lower semicontinuous. Further, z is constant in a C^1-neighbourhood of any C^1-function ϕ with only simple zeros. These trivial facts will become important later on.

The parabolic equation (0.1), (0.2) generates a semiflow $S(t)u_0 = u(t) = u(t, \cdot)$ on $X^1 \subset H_0^1 \subset C^0(I)$, thus $z(u(t))$ is well defined along solutions.

LEMMA 1.1. [6, 8]. Let $f(x, 0) = 0$ for all $x \in I$. Then the zero number $z(u(t, \cdot))$ is nonincreasing as a function of t along solutions $u(t, \cdot)$ of (0.1), (0.2).

Proof. With $a(t, x) := (f(x, u(t, x)))/(u(t, x))$ we write (0.1) as

$$u_t = u_{xx} + au, \tag{1.2}$$

where a is C^0. We apply the maximum principle to (1.2) to prove: if $x_1', x_2' \in I$ are such that $u(t, x_1') < 0 < u(t, x_2')$ then there exist continuous paths γ_i in $I \times [0, t]$ connecting (t, x_i') to a point $(0, x_i)$, such that $u < 0$ (resp. $u > 0$) along γ_1 (resp. γ_2). To see that assume $0 < x_1' < x_2' < 1$, the case $x_1' > x_2'$ is analogous. Let D_i be the path connected component of (t, x_i') in the relatively open set

$$K_i := \{(\tau, \xi) \in [0, t] \times I \mid (-1)^i u(\tau, \xi) > 0\}.$$

We claim that we can find elements $(0, x_i) \in D_i$. Otherwise, e.g. D_2 is contained in the strip $(0, t] \times I$. Replacing u by $ue^{\alpha t}$ does not change d_2 and allows us to assume $a < 0$, hence $Au := u_{xx} - u_t \geqq 0$ on $\bar{D}_2$. Let $M := \max_{\bar{D}_2} u > 0$ and choose a point $(\bar{t}, \bar{x}) \in \bar{D}_2$ with minimal $\bar{t}$ such that $u(\bar{t}, \bar{x}) = M$. From $M > 0$ we conclude $(\bar{t}, \bar{x}) \in D_2$, hence $\bar{t} > 0$. This implies a contradiction to the strong maximum principle: let $E := D_2$ and apply [7, III.2, lemma 3] to conclude $u < M$ on $d_2 \cap (\{\bar{t}\} \times I)$, contradicting $(\bar{t}, \bar{x}) \in D_2 \cap (\{\bar{t}\} \times I)$. Therefore there are points $(0, x_i) \in D_i$.

Invoking the Jordan curve theorem comletes the proof. ■

As a trivial but illustrative application, we prove estimate (0.8) for finite linear combinations

$$\phi^0 = \sum_{k=i}^{j} \alpha_k \cdot \phi_k \tag{1.3}$$

of Sturm–Liouville eigenfunctions ϕ_k for the potential $a(x) := f_u(x, v(x))$. We use the flow (1.2), defining a solution $\phi(t, \cdot)$ with initial condition $\phi(0, \cdot) = \phi^0$ and Dirichlet conditions.

COROLLARY 1.2. If the Sturm–Liouville potential a is continuous, $0 \leqq i < j < \infty$ and $\phi^0 \not\equiv 0$, then

$$i \leqq z(\phi^0) \leqq j$$

Proof. We use the explicit representation

$$\phi(t, \cdot) = \sum_{k=i}^{j} \alpha_k e^{\lambda_k t} \phi_k \tag{1.4}$$

of the solution $\phi(t, \cdot)$, $t \in \mathbb{R}$ of (1.2) through ϕ^0. From (1.4), $\phi^0 \not\equiv 0$ it is immediate that there exist integers $k^{\pm} \in \{i, i+1, \ldots, j\}$ such that

$$\lim_{t \to \pm\infty} \phi(t)/|\phi(t)| = \operatorname{sign}(\alpha_{k^{\pm}})\phi_{k^{\pm}}$$

in the C^1-topology (normalizing $|\phi_k| = 1$), because the λ_k are pairwise disjoint. The ϕ_k have

only simple zeros, hence z is constant in a C^1-neighbourhood of ϕ_k. By monotonicity of z along solutions of (1.2) (lemma 1.1) we conclude for $T > 0$ sufficiently large

$$i \leqq k^+ = z(\phi(T, \cdot)/|\phi(T, \cdot)|) = z(\phi(T, \cdot)) \leqq z(\phi^0) \leqq z(\phi(-T, \cdot)) = k^- \leqq j$$

and the proof is complete. ■

Note that the corollary holds even if $j = \infty$.

2. ZEROS ON THE UNSTABLE MANIFOLD

In this section we prove that for any element w of an n-dimensional unstable manifold of v there are less than n zeros of $w–v$. On our way we investigate the fine structure of the unstable manifold. Finally we relate $z(w - v)$ to the number of zeros of v_x.

Let v be a hyperbolic stationary solution of (0.1), (0.2) with eigenvalues $\lambda_0 < \ldots < \lambda_{n-1} < 0 < \lambda_n < \ldots$ of the linearization (0.6), (0.7) and eigenfunctions ϕ_k. By E^s, E^u, $E^s \oplus E^u = I$ we denote the complementary projections of X onto the stable and unstable spaces of the linearization $L = A - F'(v)$ at v, and by E_k, $k = 0, \ldots, n-1$, $E_0 \oplus E_1 + \ldots \oplus E_{n-1} = E^u$ the projections onto the subspaces spanned by ϕ_k.

THEOREM 2.1. Let v be a hyperbolic stationary solution as above. Then there exists an increasing sequence $W_0 \subset \ldots \subset W_{n-1} = W^u$ of invariant C^κ-submanifolds of the unstable manifold W^u through v such that

(i) $\dim W_k = k + 1$, and the tangent space to W_k at v is spanned by $\phi_0, \ldots, \phi_k$;

(ii) for any $w \in W_k \setminus W_{k-1}$

$$\lim_{t \to -\infty} (S_t(w) - v)/|S_t(w) - v| = \pm \phi_k \tag{2.1}$$

where the flow S_t for $t < 0$ is defined by $S_{-t}(S_t(w)) = w$ on W^u;

(iii) for $w \in W_k \setminus W_{k-1}$ and t near $-\infty$ the zero number z satisfies

$$z(S_t(w) - v) = k;$$

and $S_t(x) - v$ has precisely k simple zeros in (0,1);

(iv) for $w \in W_k \setminus W_{k-1}$, we obtain

$$z(w - v) \leqq k,$$

and consequently for all $w \in W^u$

$$z(w - v) < \dim W^u.$$

Note that by [5, Section 7.3], S_t is well defined for $t < 0$ on W^u.

At the end of this section we outline a simple idea for the proof of theorem 2.1 which uses finite dimensionality of W^u. Another idea which also works for the infinite dimensional stable manifold (see Section 3) can be illustrated in the case $\dim W^u = 2$. The linearization of the flow on W^u near v looks like Fig. 1, where ϕ_0, ϕ_1 are represented by the coordinate vectors. All integral curves $\gamma(t) = \alpha_0(t)\phi_0 + \alpha_1(t)\phi_1$ which are not identically zero have the property $\alpha_0(t)\alpha_1^{-1}(t) \to 0$ for $t \to -\infty$ except of two which have $\alpha_1(t) = 0$. Qualitatively, this picture is not destroyed by nonlinearities. The exceptional trajectories become W_0 in the notation of the

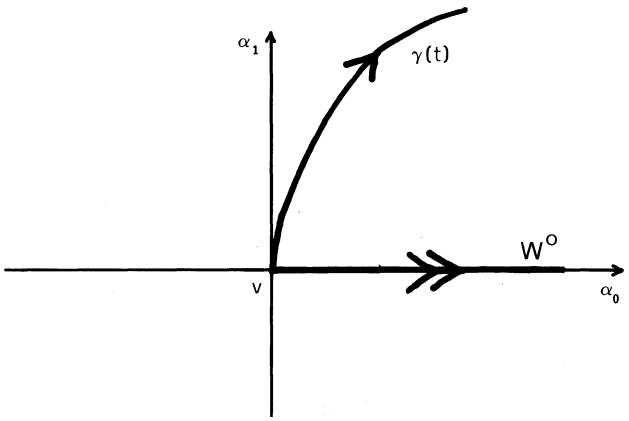


Fig. 1. The strongly unstable manifold W^0 and a general trajectory γ outside W^0.

theorem. A trajectory γ on W^u satisfies

$$\gamma(t) = v + \alpha_0(t)\phi_0 + \alpha_1(t)\phi_1 + O(|\alpha_0(t)| + |\alpha_1(t)|) \quad \text{for } t \to -\infty.$$

The exceptional ones satisfy in addition $\alpha_1(t) = o(\alpha_0(t))$, all the others $\alpha_0(t) = o(\alpha_1(t))$ for $t \to -\infty$. Consequently, $\alpha_0^{-1}(t)(\gamma(t) - v)$ mimicks ϕ_0 in the first case while $\alpha_1^{-1}(t)(\gamma(t) - v)$ mimicks ϕ_1 in the second case for t near $-\infty$. In particular, it will have the same zero number as ϕ_0, ϕ_1 respectively. We employ lemma 1.1 to conclude that $(\gamma(t) - v)$ does not increase with t, hence

$$z(\gamma(0)) \leqq \max(z(\phi_0), z(\phi_1)) = 1.$$

To carry out the idea in detail we need the following.

LEMMA 2.2. Consider a differential equation on a neighbourhood U of the origin in $\mathbb{R}^n = \mathbb{R}^p \times \mathbb{R}^q$ defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{y}) \tag{2.2}$$

$$\dot{\mathbf{y}} = \mathbf{B}\mathbf{y} + \mathbf{g}(\mathbf{x}, \mathbf{y}) \tag{2.3}$$

($\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^q$). Assume that all eigenvalues of $\mathbf{A}$ ($\mathbf{B}$) have negative real parts $\leqq a_0$ ($\geqq b_0$, respectively), where $a_0 < b_0 < 0$, $\mathbf{f}$, $\mathbf{g}$ are C^k, $k > 0$ and satisfy

$$\lim_{(\mathbf{x},\mathbf{y})\to 0} \mathbf{f}(\mathbf{x}, \mathbf{y})\,|(\mathbf{x}, \mathbf{y})|^{-1} = \mathbf{0}, \quad \lim_{(\mathbf{x},\mathbf{y})\to 0} \mathbf{g}(\mathbf{x}, \mathbf{y})\,|(\mathbf{x}, \mathbf{y})|^{-1} = \mathbf{0}.$$

Then, there exists a positively invariant neighbourhood Ω of $\mathbf{0}$ and a p-dimensional C^k submanifold W of Ω through $(\mathbf{0}, \mathbf{0})$ tangent to the subspace $\mathbf{y} = \mathbf{0}$ at $(\mathbf{0}, \mathbf{0})$ such that each solution $(\mathbf{x}(t), \mathbf{y}(t))$ of (2.2), (2.3) with $(\mathbf{x}(0), \mathbf{y}(0)) \in \Omega \backslash W$ satisfies

$$\lim_{t\to\infty} |\mathbf{y}(t)|^{-1}\,\mathbf{x}(t) = \mathbf{0}. \tag{2.4}$$

Proof. For the finite dimensional case considered here, it is easy to prove (2.4) directly from (2.2), (2.3), choosing suitable scalar products on $\mathbb{R}^p$, $\mathbb{R}^q$ and deriving a differential inequality for $\eta(t) := |\mathbf{x}(t)|^2/|\mathbf{y}(t)|^2$. However, we give a different proof which carries over without change to an infinite dimensional situation occurring in the stable manifold (see lemma 3.1 and its proof in the appendix).

The existence of Ω and an invariant manifold W tangent to the subspace $\mathbf{y} = \mathbf{0}$ at $(\mathbf{0}, \mathbf{0})$ follows from [4, lemma 4.1 and corollary 5.1, chapter IX]. If Ω is chosen sufficiently small, W can be represented as the graph of a C^k function $\mathbf{h}$ from some neighbourhood of $\mathbf{0}$ in the x-space into $\mathbb{R}^q$ with $\mathbf{h}'(\mathbf{0}) = \mathbf{0}$. It follows from [4] that if one introduces in Ω new coordinates $\mathbf{u} = \mathbf{x}$, $\mathbf{v} = \mathbf{y} - \mathbf{h}(\mathbf{x})$ then the $(\mathbf{u}, \mathbf{v})$-representation $\mathbf{\Phi}: (\mathbf{u}, \mathbf{v}) \to (\mathbf{u}_1, \mathbf{v}_1)$ of the time one map of (2.2), (2.3) satisfies

$$\mathbf{u}_1 = \tilde{\mathbf{A}}\mathbf{u} + \mathbf{U}(\mathbf{u}, \mathbf{v}) \tag{2.5}$$

$$\mathbf{v}_1 = \tilde{\mathbf{B}}\mathbf{v} + \mathbf{V}(\mathbf{u}, \mathbf{v}) \tag{2.6}$$

with $\mathbf{U}$, $\mathbf{V}$ having similar properties as $\mathbf{f}$, $\mathbf{g}$ in (2.2), (2.3) and, in addition, $\mathbf{V}(\mathbf{u}, \mathbf{0}) = \mathbf{0}$. The time one map of a differential equation maps initial values of its solutions into their values at time one.

By choosing suitable norms $|\cdot|$ in the $\mathbf{u}$-, $\mathbf{v}$-spaces we can assume

$$|\mathbf{A}\mathbf{u}| < (a + \theta)\,|\mathbf{u}|$$

$$|\mathbf{B}\mathbf{v}| > (b - \theta)\,|\mathbf{v}|$$

where $a := \exp a_0$, $b := \exp b_0$ and $0 < \theta < (b - a)/2$, $\theta < b$. Also, there is a positive function $\kappa(\rho)$ on some right neighbourhood of zero such that $\kappa(\rho) \to 0$ for $\rho \to 0$ and

$$|\mathbf{U}(\mathbf{u}, \mathbf{v})| < \kappa(\rho)\,(|\mathbf{u}| + |\mathbf{v}|),\ |\mathbf{V}(\mathbf{u}, \mathbf{v})| < \kappa(\rho)|\mathbf{v}|$$

if $|\mathbf{u}| + |\mathbf{v}| < \rho$.

Let now $(\mathbf{u}, \mathbf{v}) \in \Omega$ and let Ω be so small that $|\mathbf{u}_1| < |\mathbf{u}|$, $|\mathbf{v}_1| < |\mathbf{v}|$. Then, we have

$$\frac{|\mathbf{u}_1|}{|\mathbf{v}_1|} < \frac{(a + \theta)\,|\mathbf{u}| + \kappa(\rho)\,(|\mathbf{u}| + |\mathbf{v}|)}{(b - \theta)\,|\mathbf{v}| - \kappa(\rho)\,|\mathbf{v}|} = \frac{a + \theta + \kappa(\rho)}{b - \theta - \kappa(\rho)}\,\frac{|\mathbf{u}|}{|\mathbf{v}|} + \frac{\kappa(\rho)}{b - \theta - \kappa(\rho)}. \tag{2.7}$$

Let

$$\alpha \in \left(\frac{a + \theta}{b - \theta}, 1\right) \quad , \beta(\rho) := \frac{\kappa(\rho)}{b - \theta - \kappa(\rho)}.$$

We have $\lim_{\rho \to 0} \beta(\rho) = 0$ and there exists a $\rho_0 > 0$ such that $a + \theta + \kappa(\rho) < \alpha(b - \theta - \kappa(\rho))$ for any $\rho < \rho_0$. From (2.7) we have for $\varepsilon \in (0, 1 - \alpha)$

$$\frac{|\mathbf{u}_1|}{|\mathbf{v}_1|} < (\alpha + \varepsilon)\,\frac{|\mathbf{u}|}{|\mathbf{v}|} \quad \text{as soon as} \quad \frac{|\mathbf{u}|}{|\mathbf{v}|} > \frac{\beta(\rho)}{\varepsilon}. \tag{2.8}$$

Choose any $(\mathbf{u}_0, \mathbf{v}_0) \in \Omega$ with $\mathbf{v}_0 \neq 0$ and any $\gamma > 0$. We prove that there exists an $N > 0$ such that $|\mathbf{u}_k| < \gamma|\mathbf{v}_k|$ for all $k > N$ where $(\mathbf{u}_k, \mathbf{v}_k) = \mathbf{\Phi}^k(\mathbf{u}_0, \mathbf{v}_0)$. Indeed, assume the contrary. Since $(\mathbf{u}_k, \mathbf{v}_k) \to \mathbf{0}$, there exists an N_0 such that $|\mathbf{u}_k| + |\mathbf{v}_k| < \boldsymbol{\rho}_1 \leqq \boldsymbol{\rho}_0$ for all $k \geqq N_0$, where $\beta(\rho_1)/\varepsilon < \gamma$. From (2.8) it follows that

$$|\mathbf{u}_{k+1}| < \gamma|\mathbf{v}_{k+1}| \quad \text{as soon as} \quad |\mathbf{u}_k| + |\mathbf{v}_k| < \rho_1 \quad \text{and} \quad |\mathbf{u}_k| < \gamma|\mathbf{v}_k| \tag{2.9}$$

If $|\mathbf{u}_k| > \gamma|\mathbf{v}_k|$ for $k \geqq N_0$ then by (2.8) also

$$\gamma < \frac{|\mathbf{u}_k|}{|\mathbf{v}_k|} < (\alpha + \varepsilon)^{k-N_0} \frac{|\mathbf{u}_{N_0}|}{|\mathbf{v}_{N_0}|} \quad \text{for } k \geqq N_0$$

which is impossible. Thus, there exists an $N \geqq N_0$ for which $|\mathbf{u}_N| < \gamma|\mathbf{v}_N|$. By (2.9), we have $|\mathbf{u}_k| < \gamma|\mathbf{v}_k|$ for all $k > N$. Since γ was arbitrary, $\lim_{k\to\infty} |\mathbf{v}_k|^{-1}\mathbf{u}_k = 0$.

For the differential equation (2.2), (2.3) this means that if $(\mathbf{x}(t), \mathbf{y}(t))$ is its solution with $(\mathbf{x}(0), \mathbf{y}(0)) \in \Omega \backslash W$ (or, equivalently, $\mathbf{y}(0) \neq \mathbf{h}(\mathbf{x}(0))$), then

$$\lim_{\substack{k\to\infty \\ k \text{ integer}}} \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|} = 0. \tag{2.10}$$

We have

$$\begin{aligned} \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k)|} &= \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(k))|} \frac{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|}{|\mathbf{y}(k)|} \\ &\leqq \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|} \left(1 + \frac{|\mathbf{h}(\mathbf{x}(k))|}{|\mathbf{y}(k)|}\right) \\ &\leqq \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|} \left(1 + \frac{|\mathbf{h}(\mathbf{x}(k))|}{|\mathbf{x}(k)|} \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k)|}\right), \end{aligned} \tag{2.11}$$

or,

$$\frac{|\mathbf{x}(k)|}{|\mathbf{y}(k)|} \left(1 - \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|} \frac{|\mathbf{h}(\mathbf{x}(k))|}{|\mathbf{x}(k)|}\right) \leqq \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))|}.$$

Since $\mathbf{h}(\mathbf{x}) = o(|\mathbf{x}|)$, from (2.10), (2.11) we obtain

$$\lim_{\substack{k\to\infty \\ k \text{ integer}}} \frac{|\mathbf{x}(k)|}{|\mathbf{y}(k)|} = 0. \tag{2.12}$$

Let now $k \leqq t < k + 1$. By standard Gronwall estimates and the variation of constants formula we obtain:

$$|(\mathbf{x}(t), \mathbf{y}(t))| \leqq C|(\mathbf{x}(k), \mathbf{y}(k))| =: \rho \qquad \text{for all } k \text{ and } t \in [k, k+1)$$

with some $C \geqq 1$. Again by Gronwall and variation of constants we obtain

$$|\mathbf{x}(t)| \leqq C_1(|\mathbf{x}(k)| + \hat{\kappa}(\rho) \cdot (|\mathbf{x}(k)| + |\mathbf{y}(k)|))$$

$$|\mathbf{y}(t)| \geqq C_2(|\mathbf{y}(k)| - \hat{\kappa}(\rho) \cdot (|\mathbf{x}(k)| + |\mathbf{y}(k)|))$$

for all $k \in \mathbb{N}$, $t \in [k, k+1)$, suitable constants C_1, $C_2 > 0$ and a function $\hat{\kappa}(\rho)$ satisfying

$$\lim_{\rho\to 0} \hat{\kappa}(\rho) = 0$$

Thus we have (for all $k \in \mathbb{N}$, $t \in [k, k+1)$)

$$\frac{|\mathbf{x}(t)|}{|\mathbf{y}(t)|} \leqq \frac{C_1}{C_2} \cdot \frac{|\mathbf{x}(k)| \cdot |\mathbf{y}(k)|^{-1} \cdot (1 + \hat{\kappa}(\rho)) + \hat{\kappa}(\rho)}{1 - \hat{\kappa}(\rho)(1 + |\mathbf{x}(k)| \cdot |\mathbf{y}(k)|^{-1})}$$

and (2.12) readily implies (2.4), completing the proof of the lemma. ■

Proof of theorem 2.1. A neighbourhood V of v in W^u can be considered as an open subset Ω of $\mathbb{R}^n$, the coordinates $\mathbf{z} = (z_0, \ldots, z_{n-1})$ chosen in such a way that $z_k(w) = E_k(w - v)$ for $w \in W^u$ near v. Then, locally at v, the restriction of (0.1), (0.2) to W^u has the form

$$\dot{\mathbf{z}} = \mathbf{C}\mathbf{z} + \mathbf{q}(\mathbf{z}), \tag{2.13}$$

where $\mathbf{C} = \operatorname{diag}\{-\lambda_0, \ldots, -\lambda_{n-1}\}$, q is C^κ and $\mathbf{q}'(\mathbf{0}) = \mathbf{0}$.

Consider the associated system

$$\mathrm{d}\mathbf{z}/\mathrm{d}\tau = -\mathbf{C}\mathbf{z} - \mathbf{q}(\mathbf{z})$$

which is obtained from (2.13) by time reversal $\tau = -t$. This system satisfies the assumptions of lemma 2.2 with $\mathbf{x} = (z_0, \ldots, z_{n-2})$, $\mathbf{y} = z_{n-1}$. We denote by $\tilde{W}_{n-2}$ the submanifold W the existence of which is asserted in lemma 2.2. It is given by an equation

$$z_{n-1} = h_{n-2}(z_0, \ldots, z_{n-2}), \qquad \mathbf{z} \in \Omega$$

where h_{n-2} is C^κ and satisfies

$$h_{n-2}(\mathbf{0}) = 0. \tag{2.14}$$

By lemma 2.2, if $\mathbf{z}(t)$ is a solution of (2.13) with

$$\mathbf{z}(0) \in \Omega, z_{n-1}(0) \neq h_{n-2}(z_0, \ldots, z_{n-2}), \tag{2.15}$$

then

$$\lim_{t \to -\infty} |z^{n-1}(t)|^{-1} |(z_0(t), \ldots, z_{n-2}(t))| = 0. \tag{2.16}$$

From (2.14) it follows that

$$\lim_{t \to -\infty} |z_{n-1}(t)| \, |(z_0(t), \ldots, z_{n-2}(t)|^{-1} = 0 \tag{2.17}$$

if (2.15) does not hold. Since W^u is tangent to the unstable space of L, from (2.16), (2.17) it follows respectively

$$\lim_{t \to -\infty} |E_{n-1}(S_t(w) - v)|^{-1} \, |(I - E_{n-1})(S_t(w) - v)| = 0 \tag{2.18}$$

for $w \in V \backslash \tilde{W}_{n-2}$ and

$$\lim_{t \to -\infty} |(E_{n-1} + E^s)(S_t(w) - v)| \, |(I - E_{n-1} - E^s)(S_t(w) - v)|^{-1} = 0 \tag{2.19}$$

for $w \in \tilde{W}_{n-2}$. We define $W_{n-2} = \{S_t(\tilde{W}_{n-2}) | t \geqq 0\}$. By [5, theorem 6.1.9], W_{n-2} is an invariant submanifold of W^u. The properties (2.18), (2.19) obviously extend to $w \in W^u \backslash W_{n-2}$, W_{n-2}, respectively.

On W_{n-2}, the differential equation is again of form (2.13) with $\mathbf{C} = \operatorname{diag}\{-\lambda_0, \ldots, -\lambda_{n-2}\}$. Applying lemma 2.2 to the equation on W_{n-2} we obtain an $(n-2)$-dimensional submanifold $\tilde{W}_{n-3}$ of $\tilde{W}_{n-2}$ represented by

$$z_{n-2} = h_{n-3}(z_0, \ldots, z_{n-3})$$

with

$$h_{n-3}(\mathbf{0}) = 0 \tag{2.20}$$

such that

$$\lim_{t\to-\infty} |z_{n-2}(t)|^{-1} \, |(z_0(t), \ldots, z_{n-3}(t))| = 0 \tag{2.21}$$

for all solutions $z(t)$ with $z_{n-2}(0) \neq h_{n-2}(z_0, \ldots, z_{n-3})$. Again, we extend $\tilde{W}_{n-3}$ to an invariant submanifold of W_{n-2} by $W_{n-3} = \{S_t(\tilde{W}_{n-3}),\ t \geqq 0\}$. From (2.19) and (2.21) it follows that

$$\lim_{t\to-\infty} |E_{n-2}(S_t(w) - v)|^{-1} \cdot |(I - E_{n-2})(S_t(w) - v)| = 0$$

for $w \in W_{n-2} \backslash W_{n-3}$ while for $w \in W_{n-3}$ it follows from (2.19) and (2.20) that

$$\lim_{t\to-\infty} \left| \sum_{k=0}^{n-3} E_k(S_t(w) - v) \right|^{-1} \left| \left(I - \sum_{k=0}^{n-3} E_k(S_t(w) - v) \right| = 0.$$

In this way we may proceed further and after $n-1$ steps obtains all the $(k+1)$-dimensional manifolds W_k such that for $w \in W_k \backslash W_{k-1}$ we have

$$\lim_{t\to-\infty} |(I - E_k)(S_t(w) - v)| / |E_k(S_t(w) - v)| = 0.$$

This in turn implies for $w \in W_k \backslash W_{k-1}$ that

$$\lim_{t\to-\infty} (S_t(w) - v)/|S_t(w) - v| = \pm \phi_k. \tag{2.1}$$

Recall that the above limit is considered in $X^1 \subset C^1(I)$, and ϕ_k has only simple zeros with $z(\phi_k) = k$. By our remark preceding lemma 1.1 this implies

$$z(S_t(w) - v) = k$$

for t near $-\infty$.

Now we invoke lemma 1.1 for $z(u(t))$, $u(t) := S_t(w) - v$. Note that u satisfies an equation

$$u_t = u_{xx} + \hat{f}(x, u),$$
$$u(t, 0) = u(t, 1) = 0,$$

where $\hat{f}(x, u) := f(x, u + v(x)) - f(x, v(x))$. Hence $\hat{f}(x, 0) = 0$ and lemma 1.1 implies for t near $-\infty$

$$z(w - v) = z(u(0)) \leqq z(u(t)) = z(S_t(w) - v) = k.$$

This completes the proof of theorem 2.1. ■

From our theorem we deduce a relation between the number of changes of monotonicity of a hyperbolic stationary solution v (some "lap-number", cf. [6]) and the zero number $z(w - v)$ on the unstable manifold of v.

COROLLARY 2.3. Let v be a stationary hyperbolic solution of (0.1), (0.2), $v_x \not\equiv 0$, and let $w \in W^u$ be in its unstable manifold. Then

$$z(w - v) < z(v_x).$$

Proof. Due to theorem 2.1 it suffices to prove that $n := \dim W^u \leqq z(v_x)$.
The function $y := v_x$ solves the linearized equation

$$y_{xx} + f_u(x, v(x))y = 0.$$

On the other hand, the eigenfunction ϕ_{n-1} has $n + 1$ zeros on the closed interval $[0, 1]$. By the comparison theorem, between any two consecutive zeros of ϕ_{n-1} there has to be a zero of v_x. By $v_x \not\equiv 0$, all zeros of v_x are simple. This implies $z(v_x) \geqq n$ and the proof is complete. ■

We outline an alternate proof of theorem 2.1, (iv) which works only for W^u, as far as we know. Consider any trajectory $u(t)$ on $W^u \backslash \{v\}$ and let $y(t) := u(t)/|u(t)|$ be its projection onto the unit sphere. Then obviously

$$\lim_{t \to -\infty} E^s y(t) = 0.$$

Since W^u is finite dimensional, we may thus pick a sequence $t_k \to -\infty$ such that

$$\phi := \lim_{t_k \to -\infty} y(t_k) \tag{2.22}$$

exists in $X^1 \subset C^1(I)$. But ϕ is in the unstable eigenspace of v, hence Section 1 implies for t_k near $-\infty$

$$z(w - v) = z(u(0)) \leqslant z(u(t_k)) = z(y(t_k)) = z(\phi) < n = \dim W^u,$$

without any intermediate construction of W_k.

3. ZEROS ON THE STABLE MANIFOLD

We turn to investigate the zero number $z(w - v)$ on the stable manifold W^s of the hyperbolic stationary solution v of (0.1), (0.2), keeping the assumptions and notations of Section 2 in effect.

Similarly to the unstable case we need the following lemma on the fine structure of W^s.

LEMMA 3.1. Assuming hyperbolicity of v above and $f \in C^\kappa$, $\kappa \geqq 2$, there exists a decreasing sequence $W^s = W_n \supset W_{n+1} \supset \ldots$ of invariant C^κ-submanifolds of the stable manifold W^s through v such that

(i) the tangent space to W_k at v is spanned by $\phi_k, \phi_{k+1}, \ldots$

(ii) for any $w \in W_k \backslash W_{k+1}$

$$\lim_{t \to \infty} (S_t(w) - v)/|S_t(w) - v| = \pm \phi_k. \tag{3.1}$$

We defer the proof of this lemma to the appendix.

As an immediate consequence of lemma 3.1 we can conclude for $w \in W_k \backslash W_{k+1}$, $k \geqq n$, that $u(t) := S_t(w) - v$ satisfies

$$z(w-v) \geqq \lim_{t\to\infty} z(u(t)) = \lim_{t\to\infty} z(u(t)/|u(t)|) \geqq z\left(\lim_{t\to\infty} u(t)/|u(t)|\right) \tag{3.2}$$
$$= z(\pm\phi_k) = k,$$

by lower semicontinuity of z and monotonicity of z (lemma 1.1). However, this does not imply $z \geqq n$ on all of W^s, if for example

$$\bigcap_{k\geqq n} W_k \neq \{v\}.$$

To remedy this point we use the following alternative which is proved in [1]:

(i) either $z(u(t))$ stays infinite for all $t \geqq 0$;

(ii) or $z(u(t_0)) < \infty$ for some $t_0 \geqq 0$, and $u(t)$ has only simple zeros for an open dense set of $t \in [t_0, \infty)$.

Using this fact, we will conclude below that

$$\bigcap_{k\geqq n} W_k \subset \{w | z(w-v) = \infty\} \cup \{v\}.$$

THEOREM 3.2. Let v be a hyperbolic stationary solution of (0.1), (0.2) as above. Then for $w \in W_k \subseteq W^s$, $w \neq v$ we obtain

$$z(w-v) \geqq k$$

and in particular for all $w \in W^s \backslash \{v\}$

$$z(w-v) \geqq \dim W^u.$$

Proof. With the preceding remarks it is sufficient to prove for $w \neq v$

$$z(w-v) \geqq k \qquad \text{for all } w \in W_{k+1}, \qquad k \geqq n.$$

Obviously we may assume that $z(w-v) < \infty$. Then, by [1, theorem], there exists a $t \geqq 0$ such that $u(t, \cdot) = S_t(w) - v$ has only simple zeros. Because W_{k+1} has codimension 1 in W_k we may then choose $\tilde{u} \in W_k \backslash W_{k+1}$ such that

$$z(\tilde{u}) = z(u(t))$$

(just choosing $\|u - u(t)\|_{C^1(I)}$ small enough). But by the remarks above

$$z(\tilde{u}) \geqq k,$$

thus monotonicity of z (lemma 1.1) yields

$$z(w-v) = z(u(0)) \geqq z(u(t)) = z(\tilde{u}) \geqq k$$

and we are done. ■

4. APPENDIX

We give a proof of the fine structure of the stable manifold claimed in lemma 3.1. To this end we first construct an invariant manifold corresponding to a line, splitting the spectrum of the linearization. We use a general analytic semigroup setting

$$\frac{du}{dt} + Au = f(u) \tag{4.1}$$

in a Banach space X with norm $|\cdot|$, where A is sectorial linear $X \to X$; f: $U \to X$ is C^κ, where U is a neighborhood of 0 in X^α, $\kappa \geqq 1$, $0 \leqq \alpha < 1$; $f(0) = 0$.

Let $L := A - f'(0)$ have spectrum $\sigma(L)$. By $u(t; u_0)$ we denote the solution of (4.1) with initial data $u(0; u_0) = u_0 \in X^\alpha$.

The following lemma is well known in the finite dimensional case. It replaces [4, lemma 5.1 and corollary 5.1, chapter IX] in the proof of the infinite dimensional version of lemma 2.2. Its proof is modelled in close analogy to [5, theorem 5.2.1]. Nevertheless, for the convenience of the reader we give a detailed proof.

LEMMA 4.1. Assume $\gamma > 0$ is such that $\sigma(L) = \sigma_1 \cup \sigma_2$, $\sigma_1 = \sigma(L) \cap \{\mathrm{Re}\ \lambda < \gamma\}$, $\sigma_2 = \sigma(L) \cap \{\mathrm{Re}\ \lambda > \gamma\}$ is a decomposition of $\sigma(L)$ into spectral sets. Let $X = X_1 \oplus X_2$ be the decomposition of X corresponding to the decomposition of $\sigma(L)$ and let E_1 and E_2 be the spectral projections onto X_1 and X_2 respectively, $E_1 \oplus E_2 = I$.

Then there exist $\rho > 0$, $M > 0$ and a local invariant C^κ submanifold S of the ball $\{|u|_\alpha < \rho/2M\}$ such that:

(i) S is C^κ diffeomorphic under $E_2|_S$ to an open neighborhood of 0 in $X_2^\alpha := X_2 \cap X^\alpha$;

(ii) S is tangent to X_2^α at 0;

(iii) if $|E_2 u(0)|_\alpha < \rho/2M$ and $|u(t)|_\alpha e^{\gamma t} < \rho$ for all $t \geqq 0$ then $u(0) \in S$;

(iv) if $u(0) \in S$ then

$$\sup_{t \geqq 0} |u(t)|_\alpha e^{\gamma t} < \infty.$$

Proof. Without loss of generality assume $\sigma(A) \subset \{\mathrm{Re}\ \lambda > 0\}$. By L_1, L_2 denote the restrictions of L to X_1, X_2 respectively, let $T_i(t) := \exp(-L_i t)$ be the semigroup on X_i generated by L_i and $u_i := E_i u$ the X_i-component of u. Note that $\dim X_1 < \infty$, L_1 is bounded and there exist $0 < \beta < \gamma < \delta$ such that

$$\begin{aligned} &|T_1(t)| \leqq Me^{-\beta t}, |A^\alpha T_1(t)| \leqq Me^{-\beta t} \qquad \text{for } t \leqq 0, \\ &|A^\alpha T_2(t) E_2 A^{-\alpha}| \leqq Me^{-\delta t}, |A^\alpha T_2(t)| \leqq Mt^{-\alpha} e^{-\delta t} \qquad \text{for } t \geqq 0. \end{aligned} \tag{4.2}$$

Write $g(u) := f(u) - f'(0)u$ with components $g_i := E_i g$. Then there exists a positive function k on $(0, \rho_0)$, $\rho_0 > 0$ such that $k(\rho) \to 0$ for $\rho \to 0$ and

$$|g(u^1) - g(u^2)| \leqq k(\rho) |u^1 - u^2|_\alpha$$

as soon as $|u^j|_\alpha < \rho$, $j = 1, 2$. By [5, lemma 3.3.2], $u(t)$ solves (4.1) iff $u(t)$ solves the variation of constants version of (4.1)

$$\begin{aligned} u_1(t) &= T_1(t)u_1(0) + \int_0^t T_1(t-s) g_1(u(s))\, ds \\ u_2(t) &= T_2(t)u_1(0) + \int_0^t T_2(t-s) g_2(u(s))\, ds. \end{aligned} \tag{4.1}'$$

Assuming that the solution $u(t)$ satisfies

$$|u(t)|_\alpha e^{\gamma t} \text{ is bounded as } t \to \infty, \tag{4.3}$$

we conclude that for $t \to \infty$

$$|T_1(-t)u_1(t)|_\alpha \leqq Me^{\beta t} |u_1(t)|_\alpha \to 0$$

which implies

$$u_1(0) = -\int_0^\infty T_1(-s) g_1(u(s))\, ds,$$

and, again by (4.1)', we obtain

$$u(t) = T_2(t)a + \int_0^t T_2(t-s) g_2(u(s))\, ds - \int_t^\infty T_1(t-s) g_1(u(s))\, ds \tag{4.4}$$

where $a := E_2 u(0) \in X_2$.

We show that for $\rho > 0$ sufficiently small integral equation (4.4) has a unique solution $u_a(t)$ satisfying $|u_a(t)|_\alpha e^{\gamma t} < \rho$ provided $|a|_\alpha < \rho/2M$.

Let R_ρ be the set of continuous functions u: $[0,\infty)\to X^\alpha$ such that

$$\|u(\cdot)\| := \sup_{t\geqq 0} |u(t)|_\alpha \, e^{\gamma t} \leqq \rho$$

is finite. The set R_ρ endowed with the metric generated by $\|\cdot\|$ is a complete metric space. We claim that for ρ small enough and $\|a\|_\alpha < \rho/2M$, $a\in X_2^\alpha$ the map F_a defined by

$$F_a(u(\cdot))(t) := T_2(t)a + \int_0^t T_2(t-s)g_2(u(s))\,ds - \int_t^\infty T_1(t-s)g_1(u(s))\,ds$$

is a contraction $R_\rho\to R_\rho$. Indeed

$$\begin{aligned}
\|F_a(u(\cdot))\| &\leqq \sup_{t\geqq 0} e^{\gamma t}|T_2(t)a|_\alpha + \sup_{t\geqq 0}\int_0^t e^{\gamma t}|A^\alpha T_2(t-s)|\cdot|g_2(u(s))|\,ds \\
&+ \sup_{t\geqq 0}\int_t^\infty e^{\gamma t}|A^\alpha T_1(t-s)|\cdot|g_1(u(s))|\,ds \\
&\leqq M|a|_\alpha + |E_2|\sup_{t\geqq 0}\int_0^t e^{\gamma t}M(t-s)^{-\alpha}e^{-\delta(t-s)}k(\rho)|u(s)|_\alpha\,ds \\
&+ |E_1|\sup_{t\geqq 0}\int_t^\infty e^{\gamma t}Me^{-\beta(t-s)}k(\rho)|u(s)|_\alpha\,ds \qquad (4.5)\\
&\leqq M|a|_\alpha + |E_2|M\cdot k(\rho)\int_0^\infty t^{-\alpha}e^{(\gamma-\delta)t}\,dt\cdot\|u(\cdot)\| \\
&+ |E_1|Mk(\rho)\int_0^\infty e^{(\beta-\gamma)t}\,dt\cdot\|u(\cdot)\| \\
&\leqq M\cdot|a|_\alpha + Mk(\rho)\cdot C\|u(\cdot)\|,
\end{aligned}$$

with some constant C independent of ρ. Thus, if $|a|_\alpha < \rho/2M$ and $\rho>0$ is small enough that $k(\rho)\cdot C<\rho/2M$, then F_a maps R_ρ into R_ρ. Also, repeating the same steps as in (4.5) we find

$$\|F_a(u^1(\cdot)) - F_a(u^2(\cdot))\| \leqq \tfrac{1}{2}\|u^1(\cdot) - u^2(\cdot)\|$$

as soon as $\|u^j(\cdot)\|\leqq\rho$, $j=1,2$, so F_a is a contraction in R_ρ. Consequently, F_a has a unique fixed point $u(\cdot)\in R_\rho$ which solves (4.4).

The map $(u(\cdot),a)\to F_a(u(\cdot))$ is C^κ on $R_\rho\times(\{|a|_\alpha<\rho/2M\}\cap X_2^\alpha)$. Indeed, the map is linear in a and estimating as in (4.5) one obtains

$$\begin{aligned}
&\sup_{t\geqq 0} e^{\gamma t}|\varepsilon^{-1}(F_a(u(\cdot)+\varepsilon v(\cdot))(t) - F_a(u(\cdot))(t)) \\
&\quad - \int_0^t T_2(t-s)g_2'(u(s))v(s)\,ds + \int_t^\infty T_1(t-s)g_1'(u(s))v(s)\,ds|_\alpha \to 0 \qquad \text{for } \varepsilon\to 0.
\end{aligned}$$

Therefore

$$(v(\cdot),b)\to T_2(t)b + \int_0^t T_2(t-s)g_2'(u(s))v(s)\,ds - \int_t^\infty T_1(t-s)g_1'(u(s))v(s)\,ds \qquad (4.6)$$

is the Gâteaux differential of the map $(u(\cdot),a)\to F_a(u(\cdot))$. Since the map (4.6) is continuous in $(v(\cdot),b)$, the differential is Fréchet and $(u(\cdot),a)\to F_a(u(\cdot))$ is C^1. To obtain C^κ we iterate the arguments above.

By [5, 1.2.6] the fixed point $u_a(\cdot)$ of F_a is a C^κ-function of a in $\{|a|_\alpha<\rho/2M\}\cap X_2^\alpha$. Consequently the map h: $\{|a|_\alpha<\rho/2M\}\cap X_2^\alpha\to X_\alpha$ defined by

$$h(a) := u_a(0) = a - \int_0^\infty T_1(-s)g_1(u_a(s))\,ds$$

is C^κ and, since $E_2h(a) = E_2a = a$, has a C^κ inverse on its image S. Thus,

$$h\colon \{|a|_\alpha<\rho/2M\}\cap X_2^\alpha\to X_\alpha$$

is a C^κ-diffeomorphism. This proves (i) and, using $g_1'(0)=0$, as a direct consequence (ii). By definition of R_ρ, (iv) holds.

By construction and (4.4), S is invariant with respect to the semiflow (4.1). If $|E_2u(0)|_\alpha<\rho/2M$ and $|u(t)|_\alpha e^{\gamma t}<\rho$

for all $t \geqq 0$, then we have shown that $u(\cdot)$ satisfies (4.4). Since $u(\cdot) \in R_\rho$ and $u(t) = u_a(t)$ with $a := E_2(0$ $u(0) \in S$. Thus (iii) holds and the proof is complete. ■

Proof of lemma 3.1. Existence of the manifolds W_k as claimed in lemma 3.1 follows from lemma 4.1, c with $\lambda_{k-1} < \gamma < \lambda_k$.

Using existence of the manifolds W_k, we apply the proof of lemma 2.2 successively for each k on a neighborhood U of $v := 0$ (w.l.o.g.) in W_k, with coordinates $y = E_k u$ and $x = \sum_{j>k} E_j u$ as in the notation of Section 2. Note that the proof of lemma 2.2 carries over to analytic semigroups without the assumption that x is finite dimensional. Now lemma 2.2, together with $u(t) = S_t(w) \to 0$ and lemma 4.1, (ii) imply

$$\pm \phi_k = \lim_{t\to\infty} \frac{\sum_{j\geqq k} E_j u(t)}{|E_k u|} = \lim_{t\to\infty} \frac{\sum_{j\geqq k} E_j u(t)}{\left|\sum_{j\geqq k} E_j u(t)\right|} = \lim_{t\to\infty} \frac{u(t)}{|u(t)|}$$

and the proof is complete. ■

Acknowledgements—The authors are indebted to W. Alt for helpful criticism.

REFERENCES

0. Atkinson F. V., *Discrete and Continuous Boundary Problems*, Academic Press (1964).
1. Brunovský P. & Fiedler B., Simplicity of zeros in scalar parabolic equations, preprint (1984).
2. Hale J. K., *Topics in Dynamic Bifurcation Theory*, CBMS Regional Series in Mathematics No. 47, Am. Math. Soc., Providence, RI (1981).
3. Hale J. K. & do Nascimento A. S., Orbit connections in a parabolic equation, preprint.
4. Hartman P., *Ordinary Differential Equations*, 2nd edition, Birkhäuser, Boston (1982).
5. Henry D., Geometric theory of semilinear parabolic equations, *Lecture Notes in Mathematics* **840**, Springer, Berlin (1981).
6. Matano H., Nonincrease of the lap number of a solution for a one dimensional semilinear parabolic equation, *Pub. Fac. Sci. Univ. Tokyo* Sec. 1A, **29**, 401–441 (1982).
7. Protter M. & Weinberger H., *Maximum Principles in Differential Equations*, Prentice Hall, Englewood Cliffs, NJ (1967).
8. Redheffer R. M. & Walter W., The total variation of solutions of parabolic differential equations and a maximum principle in unbounded domains, *Math. Annln* **209**, 57–67 (1974).
9. Smoller J., *Shock Waves and Reaction Diffusion Equations*, Springer, New York (1983).

P. Brunovský

The attractor of the scalar reaction diffusion equation is a smooth graph

J. Dynam. Differential Equations 2(3) (1990), 293–323.

Journal of Dynamics and Differential Equations, Vol. 2, No. 3, 1990

The Attractor of the Scalar Reaction Diffusion Equation Is a Smooth Graph

Pavol Brunovský[1]

For the scalar reaction diffusion equation with Dirichlet boundary conditions, it is proved that its maximal compact attractor is the graph of a C^1 function from a subset with nonempty interior of a subspace of the state space the dimension of which is equal to the maximal Morse index of the equilibria of the equation.

KEY WORDS: Attractor; inertial manifold; zero number; reaction diffusion equation.
AMS (MOS) SUBJECT CLASSIFICATIONS: primary 35B40, secondary 34C30.

1. INTRODUCTION

Consider the scalar reaction diffusion equation

$$u_t = u_{xx} + f(u) \tag{1.1}$$

with Dirichlet boundary conditions

$$u(t, 0) = u(t, 1) = 0. \tag{1.2}$$

We assume that f is C^2 and satisfies

$$\limsup_{|s| \to \infty} s^{-1} f(s) < \pi^2$$

and that all stationary solutions of (1.1) are hyperbolic. The [generic (Brunovský and Chow, 1984)] set of such f's we denote by $\mathcal{G}$.

[1] Institute of Applied Mathematics, Comenius University, Mlynská dolina, 842 15 Bratislava, Czechoslovakia.

1040-7294/90/0700-0293$06.00/0

865/2/3-4

For $f \in \mathscr{G}$, (1.1), (1.2) can be considered as an abstract differential equation on the Hilbert space $X = L_2(0, 1)$ which generates a C^2 semiflow S_t on any of its dense subspaces X^α, $0 < \alpha \leqslant 1$ (Henry, 1981; Miklavčič, 1985), where X^α is the fractional space associated with the operator A given by $Au(x) = -u''(x)$ if defined and if $u(0) = u(1) = 0$. We note that $X^{1/2} = H_0^1$ and $X^1 = D(A) = H_0^1 \cap H^2$. The semiflow S_t is dissipative (i.e., there is a bounded set $B \subset X^\alpha$ such that each trajectory eventually enters B) and every trajectory has a compact closure (Hale *et al.*, 1984; Hale, 1987). The set E of equilibria is finite and contained in B.

By Hale *et al.* (1984) and (1987), S_t admits a maximal compact invariant set $\mathscr{A}$ which is given by

$$\mathscr{A} = \bigcup_{v \in E} W^u(v) \tag{1.3}$$

where $W^u(v)$ is the unstable manifold of v (Henry, 1981).

Brunovský and Fielder (1988, 1989) present a complete description of the connections between stationary solutions. The purpose of this paper is to prove additional properties of $\mathscr{A}$ announced by Brunovský (1989). In order to be able to formulate them, we introduce some notation.

Let $v \in E$. By $\lambda_0(v) < \lambda_1(v) < \lambda_2(v) < \cdots$ and $\phi_0(v)$, $\phi_1(v)$, $\phi_2(v)$, we denote, respectively, the eigenvalues and normalized (in X) eigenvectors of the linearization of (1.1), (1.2) at v which is the Sturm–Liouville problem

$$y'' + [f'(v(x)) + \lambda]\, y = 0, \qquad y(0) = y(1) = 0. \tag{1.4}$$

Further, for $0 \leqslant m \leqslant n$ we denote $X_m^n(v) = \operatorname{span}\{\phi_m(v), \ldots, \phi_{n-1}(v)\}$, $X_n(v) = \operatorname{span}\{\phi_n, \ldots\}$, $X_n^\alpha(v) := X_n(v) \cap X^\alpha$. The (Morse) instability index $i(v)$ of v is given by $\lambda_{i(v)-1} < 0 < \lambda_{i(v)}$ (note that since, by assumption, v is hyperbolic, $\lambda_n \neq 0$ for any $n \geqslant 0$).

The main result of this paper is the following

1.1. Theorem

Let $f \in \mathscr{G}$ and let $N := \max\{i(w): w \in E\}$. Then, given $v \in E$, there exists an open subset U of $X_0^N(v)$ and a C^1 function $h: U \to X_N(v)$ such that $\mathscr{N} :=$ graph h contains $\mathscr{A}$ and is positively and locally negatively invariant.

This theorem extends a result due to Jolly (1989) by which, for special f [of the Chafee–Infante type (Henry, 1981)], $\mathscr{A}$ is the graph of a Lipschitz continuous map. Also, for $v \in E$ with $i(v) = N$ it answers positively the conjecture of Fusco [proved for finite dimensional approximations of S_t by Fusco (1987)] according to which $W^u(v)$ is a graph over a subset of

$X_0^{i(v)}(v)$. In fact [cf. Remark 3.4(1)], the same proof can be used to establish the correctness of Fusco's conjecture also for $v \in E$ with $i(v) < N$.

Theorem 1.1 is proved in Section 3. In Section 2 an invariant manifold theorem is established which is needed in the proof of Theorem 1.1. Some technical parts of the proofs of the results of Section 2 are presented separately in the Appendix.

2. LOCAL INVARIANT MANIFOLDS CONTAINING GIVEN TRAJECTORIES

In this section we consider an abstract differential equation,

$$dy/dt = Ay + \tilde{F}(y) \tag{2.1}$$

on $Y = \mathbb{R}^n$, where $\tilde{F}$ is C^1 on some neighborhood of 0 and satisfies

$$\tilde{F}(0) = 0, \qquad D\tilde{F}(0) = 0.$$

Equation (2.1) generates a local C^1 flow we denote by $\tilde{\varphi}_t$; by modifying $\tilde{\varphi}$ outside some neighborhood of 0 we can make $\tilde{\varphi}$ global.

We assume that the spectrum of A, $\sigma(A)$, is disjoint from the imaginary axis and the line $\operatorname{Re} \lambda = -\beta$. Then we have

$$\sigma(A) = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$$

where

$$\Lambda_1 = \{\lambda \in \sigma(A) \colon \operatorname{Re} \lambda > 0\},$$

$$\Lambda_2 = \{\lambda \in \sigma(A) \colon 0 > \operatorname{Re} \lambda > -\beta\},$$

$$\Lambda_3 = \{\lambda \in \sigma(A) \colon \operatorname{Re} \lambda < -\beta\}.$$

By P_i and Y_i, $i = 1, 2, 3$, we denote the spectral projection corresponding to Λ_i and its image, respectively, and we write $A_i := A|_{Y_i}$. For $y \in Y$ we write $y_i = P_i y$, $\tilde{F}_i(y) = P_i \tilde{F}(y)$, $i = 1, 2, 3$. Adopting these notations we can write (2.1) equivalently as

$$y_i + A_i y_i = \tilde{F}_i(y_1, y_2, y_3), \qquad i = 1, 2, 3. \tag{2.2}$$

It is well known that a scalar product $\langle \cdot, \cdot \rangle$ can be chosen in Y in such a way that the projections P_i are orthogonal and that for suitable $\delta > 0$, $0 < \gamma < \beta - \delta$, we have

$$\langle y_1, A_1 y_1 \rangle \geqslant \gamma |y_1|^2 \tag{2.3}$$

$$-(\beta - \delta)|y_2|^2 \leqslant \langle y_2, A_2 y_2 \rangle \leqslant -\gamma |y_2|^2 \tag{2.4}$$

$$\langle y_3, A_3 y_3 \rangle \leqslant -(\beta + \delta)|y_3|^2 \tag{2.5}$$

where the norm $|\cdot|$ is generated by this scalar product. This follows, e.g., from Palis and de Melo (1980, Corollary to Theorem 2.5, Chap. 2).

The formulation of the main proposition of this section as well as some of the arguments become more transparent after a coordinate change which places certain local invariant manifolds through 0 into coordinate planes. Those are

(i) the unstable manifold $W^u(0)$ which is C^1 and tangent to Y_1 at 0,

(ii) the stable manifold $W^s(0)$ which is C^1 and tangent to $Y_2 + Y_3$ at 0,

(iii) a locally invariant C^1 manifold V which is tangent to $Y_1 + Y_2$ at 0, and

(iv) the invariant manifold W which is tangent to Y_3 at 0.

While the existence of the unstable and stable manifolds is standard (Hartman, 1964; Palis and de Melo, 1980) and the existence of W is established, e.g., by Hartman (1964) (cf. also Brunovský and Fiedler, 1986), the existence of V does not seem to appear in this immediate form in the literature. After truncating the nonlinearity (in a way which is well known from the proofs of the center-unstable manifold theorems), it follows immediately from Chow and Lu, (1988). Alternatively, V can be obtained from general theorems establishing invariant manifolds for flows which admit splitting of the state variable into two components such that one component of the difference of two trajectories has a strictly larger exponential decay rate than the other one (Kurzweil, 1970).

We note that while $W^u(0)$, $W^s(0)$, and W are uniquely defined, V is not. Neverthless, it does have to contain $W^u(0)$. Unlike $W^u(0)$, $W^s(0)$, and W, in general, it may not be smoother than C^1 no matter what the order of smoothness of $\tilde{F}$ is.

Since the manifolds $W^u(0)$, $W^s(0)$, V, W are tangent to Y_1, $Y_2 + Y_3$, $Y_1 + Y_2$, Y_3, respectively, and since $W \subset W^s(0)$, $W^u(0) \subset V$, there exists a local C^1 coordinate transformation $x = \Phi(y)$ with $\Phi(0) = 0$, $D\Phi(0) = \mathrm{id}$, which places $W^u(0)$, $W^s(0)$, V, W into Y_1, $Y_2 + Y_3$, $Y_1 + Y_2$, Y_3, respectively. We work in this new coordinate system. Because of the lack of higher smoothness of Φ, some care is needed, however. Although Φ conjugates $\tilde{\varphi}_t$ with a C^1 flow φ_t in the x-space (by $\varphi_t = \Phi \circ \tilde{\varphi}_t \circ \Phi^{-1}$), the vector field $x \mapsto Ax + F(x)$, $F(x) = D\Phi(\Phi^{-1}(x))\, F(\Phi^{-1}(x))$ that generates φ_t may not be smoother than C^0 any more (Palis and de Melo, 1980). For this reason we have to avoid the differential equation in some of our arguments and work directly with the flow instead. This complication turns

out to be minor and outweighed by better transparence of the statement and arguments in the new coordinates.

Since $D\Phi(0)$ is the identity we have

$$D\varphi_t(0) = D\tilde{\varphi}_t(0) = e^{At} \qquad \text{for all } t.$$

Hence, we have

$$\varphi_t(x) = e^{tA}x + R(t, x) \tag{2.6}$$

where R is C^1,

$$R(t, 0) = 0 \qquad \text{and} \qquad |D_x R(t, x)| \leqslant L(|x|) \tag{2.7}$$

for $0 \leqslant t \leqslant 1$ and $|x|$ sufficiently small, $L: \mathbb{R}^+ \to \mathbb{R}^+$ satisfying

$$\lim_{\eta \to 0} L(\eta) = 0. \tag{2.8}$$

In addition, since φ_t leaves Y_1, $Y_1 + Y_2$, $Y_2 + Y_3$, and Y_3 invariant, we have

$$R_2(t, x_1, 0, 0) = 0, \qquad R_3(t, x_1, x_2, 0) = 0, \tag{2.9}$$

$$R_1(t, 0, x_2, x_3) = 0, \qquad R_2(t, 0, 0, x_3) = 0, \tag{2.10}$$

in some neighborhood of 0, where $R_j := P_j R$ for $j = 1, 2, 3$.

Of course, when addressing the differential equation

$$\dot{x} = Ax + F(x) \tag{2.11}$$

generating φ_t, we cannot assume F is C^1 any more but we still have

$$|F(x)| \leqslant L(|x|)\,|x|, \tag{2.12}$$

with L possibly larger but still satisfying (2.8).

Denote

$$\Gamma(\eta) := \{x: |x_2| = \eta, |x_1| < \eta\},$$

$$\hat{\Gamma}(\eta) := \{x \in \Gamma(\eta): |x_3| \leqslant \eta\},$$

$$\Gamma_{12}(\eta) := (P_1 + P_2)\,\Gamma(\eta),$$

$$\Omega(\eta) = \{x: |x_1| < \eta, |x_2| < \eta\}.$$

For a given subset Σ of Y denote

$$\Phi(\Sigma) := \{\varphi_t(x): t > 0, x \in \Sigma\},$$

$$\Phi_\eta(\Sigma) := \{\varphi_t(x): x \in \Sigma, t > 0, \varphi_s(x) \in \Omega(\eta) \text{ for } 0 < s \leqslant t\}.$$

Below we frequently deal with manifolds which are positively invariant and locally negatively invariant. We call them briefly PLN-invariant.

2.1. Proposition

Let $\mathscr{R}$ be a PLN-invariant manifold for φ_t of dimension $\dim(Y_1 + Y_2)$. Assume the following for some $\eta > 0$ sufficiently small:

(i) $U := (P_1 + P_2)\mathscr{R} \cap \Gamma(\eta)$ *is an open subset of* $\Gamma_{12}(\eta)$, $U \cap Y_2 \neq \varnothing$.

(ii) *There is an open neighborhood B of $\bar{U}$ in $Y_1 + Y_2$ and a C^1 function $\sigma: B \to Y_3$ such that $\Sigma := \mathscr{R} \cap \Gamma(\eta) = \operatorname{graph} \sigma|_U$ and*

$$|\sigma(x_1, x_2)| \leqslant \eta, \tag{2.13}$$

$$|\sigma(x_1, x_2) - \sigma(x_1', x_2')| \leqslant |x_1 - x_1'| + |x_2 - x_2'|, \tag{2.14}$$

for (x_1, x_2), $(x_1', x_2') \in B$.

(iii) $\mathscr{R} \cap \Omega(\eta) = \Phi_\eta(\Sigma)$.

(iv) $\Phi(\Gamma(\eta_1) \cap \mathscr{R}) = \Phi(\Gamma(\eta_1)) \cap \mathscr{R}$ *for* $0 < \eta_1 \leqslant \eta$.

(v) $\mathscr{R} \cap W^u(0) = \varnothing$.

Then, $\mathscr{R}$ extends to a C^1 PLN-invariant manifold $\mathscr{M} = \mathscr{R} \cup \Phi(D)$ containing $W^u(0)$, where D is a locally invariant open disk of dimension $\dim(Y_1 + Y_2)$ such that $0 \in D$.

The proof requires several preparatory lemmas. Observe that because of Lemma A.1 (ii) we have $|P_3 x| < 2\eta$ provided $x \in \Phi_\eta(\Sigma)$ and $\eta > 0$ is sufficiently small.

2.2. Lemma

For $\eta > 0$ sufficiently small let $U \subseteq \Gamma_{12}(\eta)$ and let B be an open neighborhood of $\bar{U}$ in $Y_1 + Y_2$. Let $\sigma: B \to X_3$ be C^1 and satisfy (2.13), (2.14) for each (x_1, x_2), $(x_1', x_2') \in B$. Then, $(P_1 + P_2)\Phi_\eta(\Sigma)$, where $\Sigma := \sigma(U)$, is an open subset of $Y_1 + Y_2$ and there exists a C^1 function $s: (P_1 + P_2)\Phi_\eta(\Sigma) \to X_3$ such that

(i) $\Phi_\eta(\Sigma) = \operatorname{graph} s$;

(ii) *for each $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$|s(x_1, x_2)| \leqslant \varepsilon |x_2|, \tag{2.15}$$

$$|s'(x_1, x_2)| < \varepsilon, \tag{2.16}$$

$$|s(x_1, x_2) - s(x_1', x_2')| < \varepsilon(|x_1' - x_1'| + |x_2 - x_2'|), \tag{2.17}$$

provided $|x_2|, |x_2'| < \delta$;

(iii) *s extends to a C^1 function in a neighborhood of each point $(x_1, x_2) \in (P_1 + P_2)\,\overline{\Phi_\eta(\Sigma)} \setminus Y_1$.*

Proof

By (2.13), (2.14), and (2.11), for $x \in \bar{\Sigma}$ we have

$$\langle x_2, A_2 x_2 + F_2(x) \rangle < -\gamma\,|x_2|^2 + |x_2|\,|F_2(x)| \leqslant (-\gamma + L(\eta))\eta^2 < 0 \qquad (2.18)$$

provided η is so small that $L(\eta) < \gamma$. The inequality (2.18) means that the tangent vector to the trajectory of $x \in \Sigma$ at x is not contained in $T_x\Sigma$. Therefore, $\Phi(\Sigma)$ and $\Phi_\eta(\Sigma)$ [as an open subset of $\Phi(\Sigma)$] are C^1 submanifolds of Y and

$$\dim \Phi_\eta(\Sigma) = \dim \Phi(\Sigma) = \dim(Y_1 + Y_2). \qquad (2.19)$$

We prove that for $\eta > 0$ sufficiently small, $\Phi_\eta(\Sigma)$ is a graph over its $(P_1 + P_2)$-projection, i.e., that for any $(x_1, x_2) \in (P_1 + P_2)\,\Phi_\eta(\Sigma)$, there exists a unique $x_3 \in Y_3$ such that $(x_1, x_2, x_3) \in \Phi_\eta(\Sigma)$.

The proof is indirect. Assume that for some (x_1, x_2) there exist x_3, x_3' such that both $x := (x_1, x_2, x_3) \in \Phi_\eta(\Sigma)$ and $x' := (x_1, x_2, x_3') \in \Phi_\eta(\Sigma)$. Then there are $\xi \neq \xi' \in \Sigma$ and $t, \tau \geqslant 0$ such that $x = \varphi_{t+\tau}(\xi)$ and $x' = \varphi_t(\xi')$. By Lemma A.5, for sufficiently small $\eta > 0$,

$$|P_3(\varphi_{\tau+s}(z) - \varphi_s(z')|\;|(P_1 + P_2)(\varphi_{\tau+s}(z) - \varphi_s(z')|^{-1}$$

is bounded for $s \geqslant 0$ by some constant $c > 0$. For $s = t$ we obtain

$$0 \neq |x_3 - x_3'| \leqslant c(|x_1 - x_1'| + |x_2 - x_2'|) = 0, \qquad (2.20)$$

with $x_1 := x_1'$, $x_2 := x_2'$, a contradiction.

For $(x_1, x_2) \in (P_1 + P_2)\,\Phi_\eta(\Sigma)$ we can now define $s(x_1, x_2)$ as the unique x_3 such that $(x_1, x_2, x_3) \in \Phi_\eta(\Sigma)$. By the implicit function theorem, s is C^1 if and only if $(P_1 + P_2)|_{\phi_\eta(\Sigma)}$ is a local diffeomorphism at each $x \in \Phi_\eta(\Sigma)$. Because of (2.19) this is equivalent to

$$(P_1 + P_2)\,y \neq 0 \qquad \text{for any} \quad 0 \neq y \in T_x\Phi_\eta(\Sigma).$$

This, however, follows immediately from (2.20) if we let $x_1' \to x_1$, $x_2' \to x_2$.

To prove (ii) we first show that there exists a function $T\colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $T(\delta) \to \infty$ for $\delta \to 0$ and $t \geqslant T(\delta)$ as soon as $x \in \Phi_\eta(\Sigma)$, $|x_2| < \delta$, $x \in \varphi_t(\Sigma)$.

Indeed, assume that this is not the case. Then there exists a sequence of points $x^k \in \Phi_\eta(\Sigma)$ such that $x_2^k \to 0$ and $x^k = \varphi_{t_k}(\xi^k)$, $t_k \to t^* < \infty$, $\xi^k \to \xi^* \in \bar{\Sigma}$. By continuity we have $P_2(\varphi_{t^*}(x^*) = 0$, which contradicts Corollary A.2.

Define $k(t)$ and $\Delta(t)$ by (A.6) and (A.35), respectively. From Lemma A.1 (iv), and Lemma A.5 it follows, respectively, that there exists a $T > 0$ such that

$$k(t) < \varepsilon, \qquad \Delta(t) < \varepsilon \qquad \text{for} \quad t \geqslant T. \tag{2.21}$$

Let $\delta > 0$ be so small that $T(\delta) \geqslant T$. Then (2.21) means that (2.15) and (2.17) are satisfied for $|x_2|, |x_2'| < \delta$; (2.16) follows immediately from (2.17) for $|x_j - x_j'| \to 0$ for $j = 1, 2$. This completes the proof of (ii).

To prove (iii) we first show

$$\overline{\Phi_\eta(\Sigma)} \setminus Y_1 \subseteq \Phi_{2\eta}(\bar{\Sigma}). \tag{2.22}$$

Indeed, let $\{x^n\} \to x$, $x^n = \varphi_{t_n}(\xi^n)$, $\xi^n \in \Sigma$, and $x \notin Y_1$, i.e., $x_2 \neq 0$. Then $\{t_n\}$ is bounded by Lemma A.1 (iii) and, therefore, we may assume $\xi^n \to \xi \in \bar{\Sigma}$, $t_n \to t < \infty$. By continuity of φ_t we have $\varphi_t(\xi) = x$, $|(P_1 + P_2)\, \varphi_s(\xi)| < 2\eta$ for $0 < s \leqslant t$, hence $x \in \Phi_{2\eta}(\bar{\Sigma})$.

Because of (2.20), to obtain an extension of s to a neighborhood of $(P_1 + P_2)x$, we repeat its construction with U replaced by some neighborhood of $\bar{U}$ in the sphere $|x_2| = \eta$ and $\Omega(\eta)$ replaced by $\{x: |x_1| < 2\eta, |x_2| < \eta\}$. ∎

Let now U, η, σ, and s be as in Lemma 2.2. Extend s to $\overline{(P_1 + P_2)\, \Phi_\eta(\Sigma)} \cap Y_1$ by defining $s(x_1, 0) = 0$ for $(x_1, 0) \in \overline{(P_1 + P_2)\, \Phi_\eta(\Sigma)}$. By Lemma 2.2 (iii), at each point $(x_1, x_2) \in \overline{(P_1 + P_2)\, \Phi_\eta(\Sigma)} \setminus Y_1$, s is a restriction of a C^1 function defined in some neighborhood of (x_1, x_2). Therefore, s satisfies the hypotheses of the Whitney C^1 extension theorem (Abraham and Robbin, 1967) at each such (x_1, x_2). The estimates (2.15)–(2.17) of Lemma 2.2 mean that these hypotheses are satisfied at points of $\overline{(P_1 + P_2)\, \Phi_\eta(\Sigma)} \cap Y_1$ as well with 0 as the candidate for $s'(x_1, 0)$. Applying the Whitney extension theorem we obtain the following.

2.3. Corollary

Let the assumptions of Lemma 2.2 be satisfied. Then s extends to a C^1 function $\tilde{s}$ on $Y_1 + Y_2$ such that $\tilde{s}(x_1, 0) = 0$, $\tilde{s}'(x_1, 0) = 0$ if $(x_1, 0) \in (P_1 + P_2)\, \Phi_\eta(\Sigma) \cap Y_1$.

Combining Lemma 2.2 and Corollary 2.3 we obtain the following.

2.4. Lemma

Let Σ, η, σ, U, B, and $\tilde{s}$ be as in Lemma 2.2 and Corollary 2.3. Assume $(Y_2+Y_3)\cap U\neq\emptyset$. Then there exists a $0<\eta_1\leqslant\eta$ such that $\sigma_1:=\tilde{s}|_{\Gamma_{12}(\eta_1)}$ extends to a C^1 function $s_1: Y_1+Y_2\to U_3$ such that

$$\text{graph } s_1\cap\Omega(\eta_1)=\Phi_{\eta_1}(\text{graph } \sigma_1)\cup(Y_1\cap\Omega(\eta_1)$$

is a locally invariant manifold of φ_t containing $(\Phi_\eta(\Sigma)\cup Y_1)\cap\Omega(\eta_1)$.

Proof

Since locally $Y_2+Y_3=W^s(0)$, for $x\in U\cap(Y_2+Y_3)\neq 0$ we have $\lim_{t\to\infty}\varphi_t(x)=0$, which implies $0\in\overline{\Phi_\eta(\Sigma)}$. By Corollary 2.3 we have $\tilde{s}(0,0)=0$, $\tilde{s}'(0,0)=0$. Thus, for $\eta_1\leqslant\eta$ sufficiently small the function $\sigma_1:=\tilde{s}|_{(P_1+P_2)\Gamma(\eta_1)}$ admits a C^1 extension to a neighborhood of $(P_1+P_2)\Gamma(\eta_1)$ satisfying (2.13), (2.14) with η replaced by η_1 and $U:=(P_1+P_2)\Gamma(\eta_1)$. Also, by its definition, σ_1 admits a C^1 extension to a neighborhood of $\bar{U}$. Applying Lemma 2.2 (i) to σ_1 instead of σ allows us to define s_1: $(P_1+P_2)\Phi_\eta(\Sigma_1)\to Y_3$ by graph $s_1:=\Phi_{\eta_1}(\Sigma_1)$, where $\Sigma_1:=\text{graph }\sigma_1$. Extend s_1 to $Y_1\cap\Omega_1(\eta_1)$ by defining $s(x_1,0):=0$. Near any point of $\overline{(P_1+P_2)\Phi_{\eta_1}(\Sigma_1)}\setminus Y_1$, s_1 is a restriction of a C^1 function by Lemma 2.2 (iii) (applied to σ_1), while at any point of $Y_1\cap\Omega(\eta_1)$, s_1 satisfies the assumptions of the Whitney C^1 extension theorem because of Lemma 2.2 (ii) (applied to σ_1). Therefore, s_1 extends to a C^1 function on Y_1+Y_2.

Trivially, both $\Phi_{\eta_1}(\Sigma_1)$ and $Y_1\cap\Omega(\eta_1)$ are locally invariant and their union contains $(\Phi_\eta(\Sigma)\cap\Phi_{\eta_1}(\Sigma_1))\cup(Y_1\cap\Omega(\eta_1))$. Thus, all that remains to be proved is

$$(P_1+P_2)\Phi_{\eta_1}(\Sigma_1)=(P_1+P_2)\Omega(\eta_1)\setminus Y_1. \tag{2.23}$$

Since the restriction of P_1+P_2 to $\Phi_{\eta_1}(\Sigma_1)$ is a local isomorphism, $(P_1+P_2)\Phi_{\eta_1}(\Sigma_1)$ is open in $(P_1+P_2)\Omega(\eta_1)\setminus Y_1$. To prove (2.23) we show: that it is also closed in $(P_1+P_2)\Omega(\eta_1)$.

Let $(x_1,x_2)\in(P_1+P_2)\Omega(\eta_1)$ and $x_2\neq 0$. Assume that there are sequences $\xi_k\in\Sigma_1$, $\xi_k\to\xi\in\bar{\Sigma}_1$, $t_k\geqslant 0$ such that $\varphi_{t_k}(\xi_k)\in\Phi_{\eta_1}(\Sigma_1)$ and $(P_1+P_2)\varphi_{t_k}(\xi_k)\to(x_1,x_2)$.

Since $x_2\neq 0$, from Lemma A.1 (iii) it follows that $\{t_k\}$ is bounded. Therefore, we may assume that $t_k\to t^*\geqslant 0$. By continuity we have $(x_1,x_2)=(P_1+P_2)\varphi_{t^*}(\xi)\in\overline{\Phi_{\eta_1}(\Sigma_1)}$, hence $\varphi_{t^*}(\xi)\in\Phi_{\eta_1}(\Sigma_1)$ by Lemma A.1 (ii). This completes the proof. ∎

Proof of Proposition 2.1

All the hypotheses of Lemmas 2.2 and 2.4 and Corollary 2.3 being met, define s_1 as in Lemma 2.4. Denote

$$\mathscr{M} := \mathscr{R} \cup \Phi(\text{graph } s_1 \cap \Omega(\eta_1)). \tag{2.24}$$

The PLN-invariance of $\mathscr{M}$ follows immediately from its definition and the PLN-invariance of $\mathscr{R}$. By Lemma 2.4 we have

$$\Phi(\text{graph } s_1 \cap \Omega(\eta_1)) \supseteq \Phi(Y_1 \cap \Omega(\eta_1)) = W^u(0).$$

It remains to be proved that $\mathscr{M}$ is a manifold.

Since graph $\sigma_1 \subset \Gamma(\eta_1)$, applying consequently Lemma 2.4, hypotheses (v), (iv), (iii), and Lemma 2.4 again we obtain

$$\begin{aligned}
&\mathscr{R} \cap \Phi(\text{graph } s_1 \cap \Omega(\eta_1)) \\
&\quad = \mathscr{R} \cap \Phi(\Phi_{\eta_1}(\Sigma_1)) = \mathscr{R} \cap \Phi(\Gamma(\eta_1)) \cap \Phi(\Sigma_1) \\
&\quad = \Phi(\mathscr{R} \cap \Gamma(\eta_1)) \cap \Phi(\Sigma_1) = \Phi(\mathscr{R} \cap \Omega(\eta) \cap \Gamma(\eta_1)) \cap \Phi(\Sigma_1) \\
&\quad = \Phi(\Phi_\eta(\Sigma) \cap \Gamma(\eta_1)) \cap \Phi(\Sigma_1) = \Phi(\Phi_\eta(\Sigma) \cap \Gamma(\eta_1)).
\end{aligned}$$

By (2.18), the trajectories of φ_t cross $\Gamma(\eta_1)$ transversally, hence $\Phi(\Phi_\eta(\Sigma) \cap \Gamma(\eta_1)) = \mathscr{R} \cap \Phi(\text{graph } s_1 \cap \Omega(\eta_1))$ is a submanifold of Y of dimension $\dim Y_1 + \dim Y_2$. Since both $\mathscr{R}$ and $\Phi(\text{graph } s_1 \cap \Omega(\eta_1))$ are submanifolds of Y of the same dimension, so is $\mathscr{M} = \Phi(\text{graph } s_1 \cap \Omega(\eta_1)) \cup \mathscr{R}$. ∎

2.5. Remark

When introducing the invariant manifold V which is tangent to $Y_1 + Y_2$ at 0, we have mentioned that it is not unique. Proposition 2.1 gives a method to construct additional manifolds tangent to $Y_1 + Y_2$ at 0 provided one of such manifolds is known (in our case the latter is represented by the manifold which we have placed to the $Y_1 + Y_2$-plane by our coordinate transform). If one takes $\eta > 0$ sufficiently small, defines $U := \Gamma(\eta) \cap (Y_1 + Y_2)$ in Proposition 2.1 (and Lemma 2.4), and chooses a function $\sigma: U \to Y_3$ satisfying the estimates (2.13) and (2.14), then there is a unique invariant manifold tangent to $Y_1 + Y_2$ and containing graph σ. Note that the right-hand sides of (2.13) and (2.14) can be replaced by $p\eta$ and q, respectively, $p > 0$ and $q > 0$ arbitrary.

3. PROOF OF THEOREM 1.1

Recall that we denote

$$N = \max_{w \in E} i(w); \tag{3.1}$$

by $|\cdot|$ we denote the norm of X. Also, recall the notation introduced before Theorem 1.1. We start our proof by preparatory lemmas. Their proofs are heavily dependent on the papers (Brunovský and Fiedler, 1986, 1988, 1989). Therefore, we have to introduce some notation used there.

By the zero number of a continuous function $v \not\equiv 0$ on $[0, 1]$, denoted by $z(v)$, we understand the number of its strict sign changes (Brunovský *et al.*, 1986, 1988, 1989). We denote

$$Z_n := \{v \in E: z(v) = n \text{ or } v \equiv 0\}$$

if $f(0) = 0$, n is even and $i(0) = n$ or $n+1$ and

$$Z_n := \{v \in E: z(v) = n\}$$

otherwise (Brunovský and Fiedler, 1989, pp. 6, 11). Further, for an interval $I \subseteq \mathbb{R}$, we denote

$$EI = \{v \in E: v'(0) \in I\}.$$

As Brunovský and Fiedler, (1989), we order the elements of E by their initial slope $v'(0)$ and we use freely the terminology above, below, maximal, neighbor, etc., relatively to this ordering.

3.1. Lemma

For all $v_1 \neq v_2 \in E$ one has $z(v_1 - v_2) < N$.

Proof

Without loss of generality assume

$$v_1'(0) > 0, \qquad |v_2'(0)| \leqslant v_1'(0), \tag{3.2}$$

[if $v_1'(0) < 0$ and $|v_2'(0)| \leqslant v_1'(0)$, replace $f(u)$ by $-f(-u)$]. Then by Brunovský and Fiedler (1989, Lemma 4.2) we have

$$z(v_1 - v_2) = z(v_1). \tag{3.3}$$

Further, we have

$$N - 1 \leqslant \max_{w \in E} z(w) \leqslant N. \tag{3.4}$$

This follows from Brunovský and Fiedler (1988, Lemma 5.1) or Brunovský and Fiedler (1989, Lemma 2.1), according to which for $0 \not\equiv w \in E$ we have

$$i(w) \in \{z(w), z(w)+1\} \tag{3.5}$$

Suppose now $z(v_1 - v_2) \geqslant N$. Then, from (3.3) and (3.4) it follows that

$$z(v_1) = z(v_1 - v_2) = N. \tag{3.6}$$

From (3.1) and (3.5) it follows that

$$i(v_1) = N. \tag{3.7}$$

We complete the proof by showing that the existence of v_1, v_2 such that (3.2), (3.6), and (3.7) hold simultaneously is contradictory.

We start the proof by showing

$$E(0, v_1'(0)) = \varnothing. \tag{3.8}$$

First, we prove

$$E(0, v_1'(0)) \cap Z_N = \varnothing. \tag{3.9}$$

Suppose that this is not true and denote $\bar{w}$ the maximal element of $E(0, v_1'(0)) \cap Z_N$. Then, $\bar{w}$ is the neighbor of v_1 in Z_N and $v_1'(0)\,\bar{w}'(0) > 0$. Thus, $i(\bar{w}) \neq i(v_1)$ by Brunovský and Fiedler (1988, Lemma 2.2). Since $i(\bar{w}) \in \{z(\bar{w}),\ z(\bar{w})+1\} = \{N, N+1\}$, by (3.4) we have $i(\bar{w}) > N$, which contradicts (3.1).

Knowing (3.9), from Brunovský and Fiedler [1989, Lemma 2.2(i)], we conclude $Z_k \cap E(0, v_1'(0)) = \varnothing$ also for $k < N$. Since $Z_k \cap E(0, v_1'(0) = \varnothing$ for $k > N$ by (3.4), this proves (3.8).

From (3.8) it follows $v_2 \in E[-v_1'(0), 0]$. In order to show that this is impossible we distinguish two cases:

$$\text{(a) } N \text{ odd}, \qquad \text{(b) } N \text{ even}.$$

In case (a) it follows from (3.2) that $v_1(x) := v_1(1-x)$ is the maximal element of $E(-\infty, 0) \cap Z_N$. Indeed, since $\hat{v}_1'(0) = -v_1'(0)$, if $w \in E(-v_1'(0), 0) \cap Z_N$, then $w \in E(0, v_1'(0))$, which contradicts (3.8) (cf. also Brunovský and Fiedler (1989, Lemma 2.5).

From Brunovský and Fiedler (1989, Lemma 2.4) it follows that $E(v_1'(0),\ v_1'(0)) \neq \varnothing$. Since $i(v_1) = i(v_1) = z(v_1) = z(v_1)$ by symmetry, we have $[E(-v_1'(0), 0) \cup (0, v_1'(0))] \cap Z_k = \varnothing$ for $k < N$ by Brunovský and Fiedler [1989, Lemma 2.6($i_\pm$)]. Therefore, $v_2'(0) = 0$. By Brunovský and Fiedler (1989, Lemma 2.3), this is possible only if $f(0) = 0$ and $v_2 \equiv 0$ and

can be excluded by the perturbation argument used at the end of the proof of Brunovský and Fiedler (1986, Theorem 1.5). In case (b), $-v_1'(0) \leqslant v_2'(0) \leqslant 0$ and $E(0, v_1'(0)) = \varnothing$ implies that the maximal element w of $E(-\infty, v_1'(0))$ satisfies $-v_1'(0) \leqslant w'(0) \leqslant 0$. If $w'(0) < 0$, by Brunovský and Fiedler (1989, Lemma 2.4) and (3.7) we have $z(w) = N-1$. Since $N-1$ is odd, $\hat{w}(x) := w(1-x) \in E$, $0 < w'(0) < v_1'(0)$, a contradiction to (3.8). Therefore, $w(0) = 0$ which can again be excluded by the perturbation argument mentioned in case (a). ▮

3.2. Lemma

For every $u_1 \neq u_2 \in \mathscr{A}$ one has $z(u_1 - u_2) < N$.

Proof

From (1.3) it follows that $u_1 \in W^u(v_1)$, $u_2 \in W^u(v_2)$ for some $v_1, v_2 \in E$. We distinguish two cases:

$$\text{(a)} \quad v_1 \neq v_2, \qquad \text{(b)} \quad v_1 = v_2.$$

Case (a)

Since $S_t(u_j) \to v_j$, $j = 1, 2$ and $t \to -\infty$, in $H_2 \cap H_1^0$ and since (by Brunovský and Fiedler, 1989, Lemma 3.2) $v_1 - v_2$ has simple zeros, for t near $-\infty$ we have $z(S_t(u_1) - S_t(u_2)) = z(v_1 - v_2)$. Since by Lemma 3.1 $z(v_1 - v_2) < N$ and since $z(S_t(u_1) - S_t(u_2))$ does not increase with time (Brunovský and Fiedler, 1989), $z(u_1 - u_2) < N$.

Case (b)

Denote $v := v_1 = v_2$ and $y(t, x) := S_t(u_1)(x) - S_t(u_2)(x)$. The function $y(t, x)$ solves the linear equation

$$y_t = y_{xx} + a(t, x)\, y \tag{3.10}$$

with the boundary conditions

$$y(t, 0) = y(t, 1) = 0,$$

where

$$a(t, x) = \int_0^1 f'((1-\vartheta)\, S_t(u_2)(x) + \vartheta S_t(u_1))\, d\vartheta.$$

We have

$$\lim_{t \to -\infty} y(t) = 0 \qquad \text{and} \qquad \lim_{t \to -\infty} a(t, x) = f'(v(x)) \tag{3.11}$$

uniformly in x.

Since $y(t) \not\equiv 0$, by Henry (1985), it follows from (3.11) that

$$\lim_{t \to -\infty} y(t)\,|y(t)|^{-1} = \pm\phi_j(v)$$

[$\phi_j(v)$ defined in Section 1] for some $0 \leqslant j < N$. This implies $z(y(t)) = z(\phi_j) = j < N$ for t near $-\infty$ and, since $z(y(t))$ does not increase with t, also $z(y(0)) = z(u_1 - u_2) < N$. ▮

3.3. Proposition

For any $v \in E$, $\mathscr{A}$ is the graph of a function $h\colon P_0^N(v)\mathscr{A} \to X_N(v)$.

Proof

The statement of the lemma is equivalent to: $u = u'$ whenever $u, u' \in \mathscr{A}$ and $u - u' \in X_N(v)$. Since by Atkinson (1964, Exercise 2, p. 549), $z(u-u') \geqslant N$ if $0 \neq u - u' \in X_N(v)$, we have $u = u'$ by Lemma 3.2. ▮

3.4. Remarks

(1) The argument used to prove Proposition 3.3 remains valid if X_0^N, X_N are replaced by any two subspaces Y, Z such that $\dim Y = N$, $Y \oplus Z = X$, $Y \cap Z = \{0\}$ and

$$z(u) \geqslant N \qquad \text{for all} \quad u \in Z. \tag{3.12}$$

In particular, by Atkinson (1964, Exercise 2, p. 549), (3.12) holds if Z is the subspace spanned by all the eigenfunctions except of the first N ones of any Sturm–Liouville problem and Y is any complement of Z.

(2) By a straightforward modification of the proofs of Lemma 3.2 and Proposition 3.3, one can prove that for any $v \in E$, $i(v) = n$, $W^u(v)$ is a graph of a function $h\colon P_0^n(v)\, W^u(v) \to X_n(v)$. This property of $W^u(v)$ has been conjectured and proved for finite dimensional approximations of (1.1), (1.2) by Fusco (1987).

3.5. Lemma

For each $v \in E$ there exists a $q > 0$ such that for any two points $w_1, w_2 \in \mathscr{A}$, one has

$$|P_N(v)(w_1 - w_2)| \leqslant q\,|P_0^N(v)(w_1 - w_2)|. \tag{3.13}$$

In other words, the function h of Lemma 3.3 is globally Lipschitz.

Proof

It follows from Chow and Lu (1988) and Foias *et al.* (1986) that for a given $v \in E$, there exists an $M > N$ and a C^1 M-dimensional PLN-

invariant submanifold $\mathscr{M}$ of X (called inertial manifold) such that $\mathscr{A} \subset \mathscr{M}$. For any chosen $v \in E$, the manifold $\mathscr{M}$ is a graph of a globally Lipschitz C^1 function $g: U \to X_M(v)$, where U is an open subset of $X_0^M(v)$.

Assume that q does not exist. Then, since $\mathscr{A}$ and the unit sphere in $X_0^M(v)$ are compact, there exist sequences $\{w_1^k\}$, $\{w_2^k\}$ such that $w_j^k \to w_j^*$, $w_j^* \in \mathscr{A}$ for $j=1,2$, $P_0^M(w_1^k - w_2^k)\,|P_0^M(w_1^k - w_2^k)|^{-1} \to y$, $|y|=1$ and $|P_N(x_1^k - w_2^k)|\,|P_0^N(w_1^k - w_2^k)|^{-1} \geqslant k$ (here and below in this proof we drop the argument v at the projection operators and their images).

Note that

$$\begin{aligned} P_N(w_1^k - w_2^k) &= P_N^M(w_1^k - w_2^k) + P_M(w_1^k - w_2^k) \\ &= P_N^M(w_1^k - w_2^k) + g(P_0^M w_1^k) - g(P_0^M w_2^k), \end{aligned}$$

hence

$$|P_N(w_1^k - w_2^k)| \leqslant |P_N^M(w_1^k - w_2^k)| + 1\,|P_0^M(w_1^k - w_2^k)|,$$

where 1 is the Lipschitz constant of g. Thus,

$$\begin{aligned} \frac{|P_N^M(w_1^k - w_2^k)|}{|P_0^N(w_1^k - w_2^k)|} &\geqslant \frac{|P_N(w_1^k - w_2^k)|}{|P_0^N(w_1^k - w_2^k)|} - 1\,\frac{|P_0^M(w_1^k - w_2^k)}{|P_0^N(w_1^k - w_2^k)|} \\ &\geqslant k - 1 - 1\,\frac{|P_N^M(w_1^k - w_2^k)|}{|P_0^N(w_1^k - w_2^k)|}, \end{aligned}$$

or

$$\frac{|P_N^M(w_1^k - w_2^k|}{|P_0^N(w_1^k - w_2^k)|} \geqslant \frac{1}{1+1}(k-1) \to \infty \qquad \text{for} \quad k \to \infty.$$

Consequently,

$$|P_0^N y| = \lim_{k\to\infty} \frac{|P_0^N(w_1^k - w_2^k)|}{|P_0^M(w_1^k - w_2^k)|} \leqslant \lim_{k\to\infty} \frac{|P_0^N(w_1^k - w_2^k)|}{|P_N^M(w_1^k - w_2^k)|} = 0$$

which means $y \in X_N^M$.

We have

$$\begin{aligned} &|P_0^M(w_1^k - w_2^k)|^{-1}(w_1^k - w_2^k) \\ &\quad = |P_0^M(w_1^k - w_2^k)|^{-1}[P_0^M(w_1^k - w_2^k) + g(P_0^M(w_1^k)) - g(P_0^M(w_2^k)] \\ &\quad = \left[I + \int_0^1 g'(P_0^M((1-\theta)w_1^k + \theta w_2^k)\,d\theta)\,y\right] |P_0^M(w_1^k - w_2^k)|^{-1} P_0^M(w_1^k - w_2^k) \\ &\quad = \left[I + \int_0^1 g'(P_0^M((1-\theta)w_1^* + \theta w_2^*)\,d\theta)\right] y + \psi_k \end{aligned}$$

where $\lim_{k\to\infty} \psi_k = 0$.

We have $y \in X_N^M(v)$, $\int_0^1 g'(P_0^M(1-\theta)w_1^* + \theta w_2^*)\, d\theta y \in X_M(v)$, hence

$$P_0^N\left[I + \int_0^1 g'(P_0^M((1-\theta)w_1^* + \theta w_2^*)\, d\theta\right] y = 0,$$

and therefore,

$$z\left(\left[I + \int_0^1 g'(P_0^M((1-\theta)w_1^* + \theta w_2^*)\, d\theta\right] y\right) \geq N$$

by Atkinson (1964, Exercise 2, p. 549).

The zero number is lower semicontinuous on X, hence for sufficiently large k we have

$$z(w_1^k - w_2^k) \geq N.$$

This contradicts Lemma 3.2. ∎

2.6. Corollary

For each $v \in E$ there exists a $q > 0$ such that

$$|P_N(v)\, y| \leq q\, |P_0^N(v)\, y|$$

for any $y \in T_u W^u(w)$, $u \in W^u(w)$, $w \in E$.

Indeed, since $W^u(w) \subset \mathscr{A}$, we have for any C^1 curve $\gamma\colon [0, \varepsilon_0) \to W^u(w)$, $\gamma(0) = u$, $\gamma'(0) = y \in T_u W^u(w)$

$$|P_N(y)| = \lim_{\varepsilon \to 0} 1/\varepsilon\, |P_N(\gamma(\varepsilon) - \gamma(0)|$$

$$\leq q \lim_{\varepsilon \to 0} 1/\varepsilon |P_0^N(\gamma(\varepsilon) - \gamma(0)| = q\, |P_0^N(v)\, y|.$$

Proof of Theorem 1.1

By Proposition 3.3, for any chosen $v \in E$ the set $\mathscr{A}$ is a graph of a function $h\colon P_0^N(v)\mathscr{A} \to X_N$. Also, from Corollary 3.6 it follows that h is C^1 on each $P_0^N(W^u(w))$, $w \in E$. It remains to be proved that there exists a PLN-invariant C^1 manifold $\mathscr{N}$ of dimension N containing $\mathscr{A}$. Since $\mathscr{A}$ is a compact attractor, it is obvious that $\mathscr{N}$ can be restricted in such a way that it will preserve its invariance properties, contain $\mathscr{A}$, and be a graph of a C^1 extension of h.

The manifold $\mathscr{N}$ will be constructed by induction. Let us order the equilibria into a sequenbce $w_1, \ldots, w_r$ in such a way that $i(w_j) \geq i(w_k)$ if $j < k$.

We construct a sequence $\{\mathcal{N}_j\}$ of C^1 locally invariant manifolds of dimension N such that $\mathcal{N}_{j+1}$ extends an open submanifold of $\mathcal{N}_j$ and

$$\mathcal{A}_j := \bigcup_{v \leqslant j} W^u(w_v) \subset \mathcal{N}_j.$$

Then, $\mathcal{N}_r$ will be a locally invariant manifold of dimension N containing A.

Denote

$$j_n := \max\{j: i(w_j) \geqslant n\}.$$

We define

$$\mathcal{N}_{j_N} := \bigcup_{1 \leqslant j \leqslant j_N} W^u(w_j).$$

For a given $j > j_N$ denote $n := i(w_j)$ and assume that $\mathcal{N}_{j-1}$ has been constructed to contain $\bigcup_{1 \leqslant v \leqslant j-1} W^u(w_v)$. To complete the induction step we extend an open submanifold of $\mathcal{N}_{j-1}$ containing $\mathcal{A}_{j-1}$ to a PLN-invariant manifold $\mathcal{N}_j$ containing $W^u(w_j)$.

To this end we employ Proposition 2.1. First we note that the inertial manifold theorem of Chow and Lu (1988) and Foias *et al.* (1986) allows us to reduce the extension step to one for finite dimensional systems. As mentioned in the proof of Lemma 3.5, [8, 9, 20] provide for an $M < N$-dimensional PLN-invariant manifold $\mathcal{M}$ which can be expressed by $\mathcal{M} :=$ graph g, where $g: Q \to X_M(w_j)$ is C^1 and Q is an open subset of $X_0^M(w_j)$ containing $P_0^M(w_j)(\mathcal{A})$. The semiflow induces a local flow φ_t on Q by

$$\varphi_t(u) = P_0^M(w_j)\, S_t(u + g(u)).$$

To simplify the formulations we extend φ_t to a global flow on $X_0^M(w_j)$ by modifying it outside some neighborhood of $P_0^M(w_j)\mathcal{A}$ if necessary.

Since $\mathcal{A} \subset \mathcal{M}$, in particular, we have $W^u(w_v) \subset \mathcal{M}$ for all v. Therefore, we may add $\mathcal{N}_{j-1} \subset \mathcal{M}$ to our induction hypotheses. In addition, we assume that

$$\mathcal{N}_{j-1} = \bigcup_{v < j} \Phi(D_v) \tag{3.14}$$

where D_v is a locally invariant open disk of dimension N containing a neighborhood of w_v in $W^u(w_v)$. From the construction of $\mathcal{N}_j$ it is seen immediately that it also lies in $\mathcal{M}$ and satisfies (3.14) with $j-1$ replaced by j.

We now introduce some notation to match that of Proposition 2.1. First, we note that the spectrum Λ of the operator L of the linearized

problem (1.4) for $v := w_j$ [defined by $(Ly)(x) = y''(x) + f'(v(x))\, y(x)$ for $y \in H_0^1 \cap H^2$] admits a partition $\Lambda \cup \Lambda_4$, $\Lambda = \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$, where $\Lambda_1 := \{-\lambda_0(w_j), \ldots, -\lambda_{n-1}(w_j)\}$, $\Lambda_2 := \{-\lambda_n(w_j), \ldots, -\lambda_{N-1}(w_j)\}$, $\Lambda_3 := \{-\lambda_N(w_j), \ldots, -\lambda_{M-1}(w_j)\}$, and $\Lambda_4 := \{\lambda_M, \ldots\}$. The corresponding splitting of $Y := X_0^M(w_j)$ is $Y = Y_1 \oplus Y_2 \oplus Y_3$, where $Y_1 := X_0^n(w_j)$, $Y_2 := X_n^N(w_j)$, and $Y_3 := X_N^N(w_j)$. Then for $A_i := L|_{\Lambda_i}$ (2.3)–(2.5) are satisfied with

$$\gamma < \min\{|\lambda_{n-1}(w_j)|, |\lambda_n(w_j)|\},$$
$$\beta = 1/2(\lambda_{N-1}(w_j) + \lambda_N(w_j)),$$
$$0 < \delta < 1/2(\lambda_N(w_j) - \lambda_{N-1}(w_j)).$$

As in Section 2 we introduce coordinates $x = (x_1, x_2, x_3)$ in such a way that $x(w_j) = 0$ and the manifolds $W^u(w_j)$, $W^s(w_j)$, V, and W (the latter two introduced in Section 2) locally at 0 coincide with Y_1, $Y_2 + Y_3$, $Y_1 + Y_2$, and Y_3; by P_i we denote the orthogonal projection $Y \to Y_j$, $j = 1, 2, 3$. We do not distinguish between $\mathcal{N}_{j-1}$, $\mathcal{A}_{j-1}$, φ_t, etc., and their representations in the x-coordinates. Then, φ_t is generated by the differential equation (2.11) and satisfies (2.6), with R satisfying (2.7)–(2.10). It is then sufficient to construct $\mathcal{N}_j$ as a submanifold of the x-space.

As the manifold $\mathcal{R}$ of Proposition 2.1 we take a suitable restriction of $\mathcal{N}_{j-1}$ of the form (3.14) (with D_v possibly replaced by their open subdisks) which contains $\mathcal{A}_{j-1}$. Below, we prove that $\mathcal{M}$ can be chosen to satisfy the hypotheses of Proposition 2.1. The PLN-invariant manifold $\mathbb{R}$ which is provided by Proposition 2.1 is of the form (3.14) and contains both $\mathcal{R}$ and $W^u(0)$. Trivially, $\mathcal{M} \subseteq \mathcal{M}$. Therefore, we can take it for $\mathcal{N}_j$. This completes the induction step and, thus, also the proof of the theorem.

It remains to be verified that the requirements of Proposition 2.1 can be met by a suitable choice of $\mathcal{R}$.

By Lemma 3.5 we have $\mathcal{A} = \text{graph } h$, where h: $(P_1 + P_2)\mathcal{A} \to Y_3$ is C^1 and satisfies

$$|h(x_1, x_2) - h(x_1', x_2')| \leqslant q(|x_1 - x_1'| + |x_2 - x_2'|)$$

for some $q > 0$ and any (x_1, x_2), $(x_1', x_2') \in (P_1 + P_2)\mathcal{A}$; by rescaling Y_3 we can achieve $q \leqslant 1/4$. Since $\mathcal{A}$ is compact it follows that there is a neighborhood C of $\mathcal{A}$ such that

$$|x_3 - x_3'| \leqslant (1/2)(|x_1 - x_1'| + |x_2 - x_2'|) \tag{3.15}$$

for any $x, x' \in (\mathcal{N}_{j-1} \cup \mathcal{A}) \cap C$.

The set $\mathcal{A}$, being the maximal compact attractor, is Lyapunov stable, i.e., for any neighborhood Q of $\mathcal{A}$ there is a neighborhood R of $\mathcal{A}$ such

that $\Phi(R) \subseteq Q$. In particular, by possibly restricting the disks D_v we can make $\tilde{\mathscr{R}} := \bigcup_{v<j} \Phi(D_v)$ to satisfy $\operatorname{cl} \tilde{\mathscr{R}} \subset C$. Then (3.15) holds for all $x, x' \in \mathscr{R} \cup \mathscr{A}$.

By Henry [1985, Properties (5) and (1), p. 191], $0 \in \operatorname{cl} W^u(w_v)$ for some v implies $W^u(w_v) \cap W^s(0) \neq 0$, the intersection being transversal by Henry (1985, Theorem 7). Since $W^s(0)$ coincides with Y_2 locally at 0, for sufficiently small $\eta > 0$ we have

$$W^u(w_v) \cap \Gamma(\eta) \cap Y_2 \neq \varnothing \qquad \text{if} \quad 0 \in \operatorname{cl} W^u(w_v), \tag{3.16}$$

$$W^u(w_v) \cap \overline{Q(\eta)} = \varnothing \qquad \text{if} \quad 0 \notin \operatorname{cl} W^u(w_v). \tag{3.17}$$

Since $i(w_v) \geqslant i(0)$ for $v \leqslant j$, from (1, Theorem 7) and (3.16), it follows that

$$\overline{Q(\eta)} \cap (\mathscr{A} - \mathscr{A}_{j-1}) \neq \varnothing \tag{3.18}$$

for $\eta > 0$ small.

By Henry (1985, Property (5, p. 191), $w_v \notin W^u(w_j)$ for $v < j$. Therefore, by possibly restricting D_v we can achieve

$$\bigcup_{v<j} D_v \cap (\overline{Q(\eta)} \cup W^u(w_j)) = \varnothing \tag{3.19}$$

for $\eta > 0$ sufficiently small.

Let $\eta > 0$ be so small that (3.17) and (3.19) hold. To complete the proof we distinguish two cases:

$$\text{(a)} \quad 0 \in \operatorname{cl} \mathscr{A}_{j-1}, \qquad \text{(b)} \quad 0 \notin \operatorname{cl} \mathscr{A}_{j-1}.$$

Case (a)

By (3.17) we have $\Gamma(\eta) \cap \mathscr{R} \cap Y_2 \neq \varnothing$; because of (3.16), (3.17), by possibly restricting D_v further, we can achieve that

$$\Phi(D_v) \cap \operatorname{cl} \Omega(\eta) = \varnothing \qquad \text{if} \quad 0 \notin \operatorname{cl} W^u(w_v) \tag{3.20}$$

provided $\eta > 0$ is sufficiently small. Then since (3.15) holds for all $x, x' \in \tilde{\mathscr{R}} \cup \mathscr{A}$, and since $0 \in \mathscr{A}$, we have

$$|x_3| = |x_3 - 0| \leqslant (1/2)(|x_2| + |x_3|) \qquad \text{for} \quad x \in \operatorname{cl} Q(\eta) \cap \tilde{\mathscr{R}}. \tag{3.21}$$

Take open subdisks G_v of D_v containing w_v such that $\bar{G}_v \subset D_v$, $v < j$, and denote $\mathscr{R} := \bigcup_{v<j} \Phi(G_v)$, $U := (P_1 + P_2)\mathscr{R} \cap \Gamma(\eta)$. Then there is a neighborhood B of $\bar{U}$ in $Y_1 + Y_2$ such that $B \subseteq (P_1 + P)\mathscr{R}$ and, therefore, we can define $\sigma := h|_B$. By (3.15) and (3.21), $\mathscr{R}$, U, B, and σ satisfy hypotheses (i) and (ii) of Proposition 2.1; by (3.19), hypothesis (v) is satisfied as well.

To verify hypothesis (iv) we first note that from $\Phi(\mathscr{R}) = \mathscr{R}$ it follows

$$\Phi(\Gamma(\eta_1) \cap \mathscr{R}) \subseteq \Phi(\Gamma(\eta_1)) \cap \mathscr{R}. \tag{3.22}$$

The opposite inclusion follows from the fact that S_t and, consequently, also φ, is gradient-like, i.e., there is a scalar function V on Y which decreases strictly along nonconstant trajectories (Henry, 1981).

Indeed, let $x \in \mathscr{R}$, $x = \varphi_t(\xi)$, $\xi \in \Gamma(\eta_1)$, $t \geqslant 0$. By definition of $\mathscr{R}$ we have $x = \varphi_{t'}(\xi')$ for some $t' \geqslant 0$, $\xi' \in D_\nu$, $\nu < j$; by (3.17) we have

$$W^u(w_\nu) \cap W^s(0) \neq 0. \tag{3.23}$$

Since V decreases along nonconstant trajectories, from (3.22) it follows $V(0) < V(w_\nu)$. If $\eta > 0$, D_ν are chosen sufficiently small it follows that $V(x') < V(x'')$ for all $x' \in \Gamma(\eta)$, $x'' \in D_\nu$, hence $V(\xi) \leqslant V(\xi')$, from which it follows that $t' - t \geqslant 0$. This means $\xi \in \Phi(\mathscr{R}) \cap \Gamma(\eta_1) = \mathscr{R} \cap \Gamma(\eta_1)$ and $x = \varphi_t(\xi) \in \Phi(\mathscr{R} \cap \Gamma(\eta_1))$. This completes the verification of hypothesis (iv).

It remains to verify hypothesis (iii). From (3.22) and the definition of Σ it follows that

$$\mathscr{R} \cap \Omega(\eta) = \Phi_\eta(\Sigma) \cup \Phi_\eta(\mathscr{R} \cap \Delta(\eta)), \tag{3.24}$$

where $\Delta(\eta) := \{x: |x_1| = \eta,\ |x_2| \leqslant \eta,\ |x_3| \leqslant \eta\}$. For $x \in \Delta(\eta)$, $x(t) := \varphi_t(x)$ we have

$$\begin{aligned}(1/2)\, d/dt\, |x_1|^2|_{t=0} &= \langle A_1 x_1 \cdot x_1 \rangle + \langle F_1(x), x_1 \rangle \\ &\geqslant [\gamma\eta^2 - L(\eta)]\eta > 0,\end{aligned}$$

provided $\eta > 0$ is sufficiently small. This proves $\Phi_\eta(\Delta(\eta)) = \varnothing$ and, by (3.17), verifies hypothesis (iii).

Case (b)

By assumption it follows from (3.17) and (3.18) that we can restrict the disks D_ν in such a way that, for sufficiently small $\eta > 0$, we have $\overline{Q(\eta)} \cap \tilde{\mathscr{R}} = \varnothing$. For such $\eta > 0$ we can chose $U := \Gamma(\eta)$, $s: \Gamma_{12}(\eta) \to Y_3$ arbitrarily satisfying (2.13), (2.14) and $\mathscr{R} := \tilde{\mathscr{R}} \cup \varphi_{(-\varepsilon, \infty)}(\text{graph } \sigma)$ for some $\varepsilon > 0$ sufficiently small. This choice of $\mathscr{R}$ obviously satisfies the hypotheses of Proposition 2.1 ▮

APPENDIX

In this Appendix we prove several technical lemmas which are needed in Section 2.

We consider the differential equation

$$\dot{x}_i = A_i x_i + F_i(x_i), \qquad i := 1, 2, 3 \tag{A.1}$$

on $Y = Y_1 + Y_2 + Y_3$ from Section 2 in the transformed coordinates. That is, we assume that A_i satisfy (2.3)–(2.5), F_i are continuous and satisfy (2.6), and (A.1) generates a unique flow φ_t which can be represented by

$$\varphi_t(x) = e^{tA}x + R(t, x)$$

with $x := (x_1, x_2, x_3)$ and $R := (R_1, R_2, R_3)$ being C^1 and satisfying (2.9) and (2.10). As in Section 2, by P_j and R_j we denote the orthogonal projection $Y \to Y_j$ and $P_j R$, respectively, $j = 1, 2, 3$.

Note that (2.3)–(2.5) imply

$$|e^{-A_1 t}| \leqslant e^{-\gamma t}, \tag{A.2}$$

$$|e^{A_2 t}| \leqslant e^{-\gamma t}, \qquad |e^{-A_2 t}| \leqslant e^{(\beta-\delta)t}, \tag{A.3}$$

$$|e^{A_3 t}| \leqslant e^{-(\beta+\delta)t} \tag{A.4}$$

for $t \geqslant 0$, respectively.

Recall the definitions of $\Gamma(\eta)$, $\Omega(\eta)$, and $\Gamma(\eta)$ from Section 2 and denote

$$\hat{\Omega}(\eta) := \{x \in \Omega(\eta) : |x_3| \leqslant \eta\}.$$

For given $\eta > 0$ define

$$p(t) = \sup\{|(P_2 + P_3)\,\varphi_t(x)| \; x \in \hat{\Omega}(\eta) \cup \hat{\Gamma}(\eta),\ \varphi_s(x) \in \Omega(\eta) \text{ for } 0 < s \leqslant t\}, \tag{A.5}$$

$$k(t) = \sup\left\{\frac{|P_3 \varphi_t(x)|}{|P_2 \varphi_t(x)|} : x \in \Gamma(\eta),\ \varphi_s(x) \in \Omega(\eta) \text{ for } 0 < s \leqslant t\right\}. \tag{A.6}$$

A.1. Lemma

For $\eta > 0$ sufficiently small we have the following.

(i) *If $x \in \hat{\Gamma}(\eta) \cap \hat{\Omega}(\eta)$ and $\varphi_s(x) \in \Omega(\eta)$ for $0 < s \leqslant t$, then*

$$|P_3 x(t)| \leqslant 2\eta. \tag{A.7}$$

(ii) *If $x \in \operatorname{cl} \hat{\Gamma}(\eta)$, $\varphi_s(x) \in \Omega(\eta)$ for $0 < s < t$ and $|P_1 \varphi_t(x)| = \eta$, then $|P_1 x(t+\tau)| > \eta$ for $\tau > 0$ sufficiently small.*

(iii) $\lim_{t \to \infty} p(t) = 0$.

(iv) $|k(t)| \leqslant 2$ *for* $t \geqslant 0$ *and*

$$\lim_{t \to \infty} k(t) = 0. \tag{A.8}$$

To simplify the formulations in the proofs in this Appendix, once we consider $\varphi_t(x)$ for some $t>0$ we automatically will assume that $\varphi_s(x)\in \operatorname{cl}\Omega(\eta)$ for $0<s\leqslant t$ without explicitly saying so. In other words, we restrict φ_t to the (local) flow in $\operatorname{cl}\Omega(\eta)$. Once we prove Lemma A.1(i) it allows us, in addition, to restrict φ_t to $\hat{\Omega}(\eta)\cup\hat{\Gamma}(\eta)$. This does not concern the formulations of the results, which are given in full.

Proof

(i) For $x\in\hat{\Gamma}(\eta)\cup\hat{\Omega}(\eta)$, η sufficiently small and $0\leqslant t\leqslant 1$ it follows from (2.9) and (2.10) that

$$|R_2(t,x_1,x_2,x_3)|\leqslant L(\eta)(|x_2|+|x_3|), \tag{A.9}$$

$$|R_1(t,x_1,x_2,x_3)|\leqslant L(\eta)|x_1|, \tag{A.10}$$

$$|R_3(t,x_1,x_2,x_3)|\leqslant L(\eta)|x_3|. \tag{A.11}$$

Denote $x(t):=\varphi_t(x)$. If $|x_3(t)|=\eta$, from (2.5) and (2.12), it follows that

$$\begin{aligned}(1/2)d|x_3(t)|^2/dt&=\langle x_3(t),A_3x_3(t)+F_3(x(t))\rangle\\&\leqslant -(\beta+\delta)\eta^2+L(\eta)\eta^2<0\end{aligned} \tag{A.12}$$

provided η is so small that $L(\eta)<\beta+\delta$. This proves (A.7).

(ii) If η is so small that (i) holds, by (2.3) and (2.11) we have

$$1/2\, d\,|x_1(t)|^2/dt=\langle x_1(t),Ax_1(t)+F_1(x(t))\rangle\geqslant\gamma\eta^2-\eta^2L(\eta)>0,$$

provided $\eta>0$ is so small that $L(\eta)<\gamma$. This proves (ii).

(iii) Let $\eta>0$ be so small that (i) holds. Then $|x_j(t)|\leqslant 2\eta$ for $j=1,2,3$, and by (A.9) we have

$$\begin{aligned}|x_2(t+1)|&\leqslant(e^{-\gamma}+L(\eta))\,|x_2(t)|+L(\eta)\,|x_3(t)|,\\|x_3(t+1)|&\leqslant(e^{-(\beta+\delta)}+L(\eta))\,|x_3(t)|,\end{aligned}$$

hence

$$|x_2(t+1)|+|x_3(t+1)|\leqslant(e^{-\gamma}+2L(\eta))(|x_2(t)|+|x_3(t)|). \tag{A.13}$$

Let $\eta>0$ be so small that $a:=e^{-\gamma}+2L(\eta)<1$. Applying (A.13) to $t=0,1,\ldots,n-1$, we obtain

$$|x_2(n)|+|x_3(n)|\leqslant 2a^n\eta.$$

If $n := [t]$, the integer part of t, we have

$$\begin{aligned}|x_2(t)| + |x_3(t)| &\leqslant e^{-\gamma(t-n)}\,|x_2(n)| + L(\eta)(|x_2(n)| + |x_3(n)|)\\ &\quad + e^{-(\beta+\delta)(t-n)}\,|x_3(n)| + L(\eta)\,|x_3(n)|\\ &\leqslant (1+2L(2\eta))(|x_2(n)| + |x_3(n)|) \leqslant 2(1+2L(\eta))a^n\eta\end{aligned}$$

Since the left-hand side of the inequality depends only on η, this proves (iii).

(iv) Let $\eta > 0$ be so small that (i) holds. If $x_2(t) \neq 0$, denote $\chi(t) = |x_3(t)|\,|x_2(t)|^{-1}$. If $\chi(t) \leqslant 1$, we have from (A.9) and (A.11)

$$\begin{aligned}\chi(t+1) - \frac{|x_3(t+1)|}{|x_2(t+1)|} &\leqslant \frac{[e^{-(\beta+\delta)} + L(\eta)]\,|x_3(t)|}{e^{-(\beta-\delta)}|x_2(t)| - L(\eta)[|x_2(t)| + |x_3(t)|]},\\ &\leqslant \frac{e^{-(\beta+\delta)} + L(\eta)}{e^{-(\beta-\delta)} - L(\eta)(1+\chi(t))}\,\chi(t).\end{aligned}$$

Let $\eta > 0$ be so small that $b := [e^{-(\beta+\delta)} + L(\eta)][e^{-(\beta-\delta)} - 2L(\eta)]^{-1} < 1$. Then, from $\chi(0) \leqslant 1$ we obtain by induction $\chi(n) \leqslant 1$ for $t \geqslant n \geqslant 0$ integer and, in turn, also $\chi(n) \leqslant b^n$.

Let now $n = [t]$. We have

$$\begin{aligned}|\chi(t)| = |x_3(t)|\,|x_2(t)|^{-1} &\leqslant [e^{-(\beta+\delta)(t-n)} + L(\eta)]\,|x_3(n)|\\ &\quad \times [e^{-(\beta-\delta)(t-n)}|x_2(n)| - L(\eta)(|x_2(n)| + |x_3(n)|)]^{-1}\\ &\leqslant (1+L(\eta))(1-2L(\eta))^{-1}\chi(n) \leqslant (1+L(\eta))(1-2L(\eta))^{-1}b^n.\end{aligned}$$

Since b and the right-hand side of the inequality depend on η only, this proves (iv). ∎

From the local invariance of $Y_1 \cap \Omega(\eta_1)$ and Lemma A.1 (iv) we obtain the following.

A.2. Corollary

For sufficiently small $\eta > 0$, one has $|P_2\varphi_t(x)| \neq 0$ provided $x \in \hat{\Gamma}(\eta)$ and $\varphi_s(x) \in Q(\eta)$ for $0 < s \leqslant t$.

In most of the arguments below there is no need to consider the components x_1 and x_2 separately. Therefore, in order to shorten the formulas we frequently aggregate them into one component $x_{12} := x_1 + x_2$. Correspondingly we write $Y_{12} := Y_1 + Y_2$, $A_{12} := A_1 + A_2$, $P_{12} := P_1 + P_2$, etc.

For fixed x, x' denote $y(t) := \varphi_t(x') - \varphi_t(x)$. From (2.6) it follows that

$$y_i(s+\tau) = e^{\tau A_i}\,y_i(s) + b_{i,12}(\tau, s)\,y_{12}(s) + b_{i,3}(\tau, s)\,y_3(s) \qquad \text{(A.14)}$$

for $0 \leqslant s \leqslant t$, $0 \leqslant \tau \leqslant \min\{1, t-s\}$, and $i = 12, 3$, where

$$b_{i,j}(\tau, s) = \int_0^1 D_{x_j} R_i(\tau, (1-\vartheta)\,\varphi_s(x) + \vartheta\varphi_s(x'))\, d\vartheta.$$

In the lemma below we consider $y(t)$ satisfying (A.14) with $b_{i,j}(\tau, t)$ such that

$$|b_{i,j}(\tau, t)| \leqslant L \quad \text{for} \quad i, j = 12, 3, \quad \tau \geqslant 0, \quad \text{and some } L > 0, \tag{A.15}$$

$$|b_{3,12}(\tau, t)| \leqslant \rho(t), \tag{A.16}$$

where ρ satisfies

$$\lim_{t \to \infty} \rho(t) = 0. \tag{A.17}$$

A.3. Lemma

Let $q > 0$ be given. Let $y(t) = (y_{12}(t), y_3(t))$ satisfy (A.14) with $b_{i,j}$ satisfying (A.15) and (A.16), ρ satisfying (A.17). Then for sufficiently small $L > 0$ there exists a positive function r: $[0, \infty) \to [0, \infty)$ depending on ρ only and satisfying

$$\lim_{t \to \infty} r(t) = 0 \tag{A.18}$$

such that if

$$|y_3(0)| \leqslant q\,|y_{12}(0)| \tag{A.19}$$

then

$$|y_3(t)| \leqslant r(t)\,|y_{12}(t)|. \tag{A.20}$$

Proof

Let $y(0)$ satisfy (A.19) and $y_{12}(0) \neq 0$. Denote

$$\lambda(t) := |y_3(t)|\,|y_{12}(t)|^{-1}.$$

If $\lambda(n) \leqslant q$, we have

$$\lambda(n+1) \leqslant \frac{(|e^{A_3}| + L)\,|y_3(n)| + \min\{\rho(n), L\}\,|y_{12}(n)|}{(|e^{-A_{12}}|^{-1} - L(1+q))\,|y_{12}(n)|},$$

$$\leqslant \chi\lambda(n) + \sigma_n,$$

where

$$\chi = \frac{e^{-(\beta+\delta)} + L}{e^{-(\beta-\delta)} - (1+q)L} = \frac{e^{-2\delta} + Le^{\beta-\delta}}{1-(1+q)\,Le^{\beta-\delta}}.$$

$$\sigma_n = \frac{\min\{\rho_n, L\}}{e^{-(\beta-\delta)} - L(1+q)}.$$

Let $L>0$ be so small that

$$\chi < 1 \tag{A.21}$$

and

$$\chi q + \sigma_n < q. \tag{A.22}$$

From (A.22) it follows $\lambda(n) \leqslant q$ by induction, and from (A.21) we obtain

$$\lambda(n) < \chi^n q + \sum_{j=0}^{n-1} \chi^{n-1-j}\sigma_j = \chi^n q + K \sum_{j=0}^{n-1} \chi^{n-1-j} \min\{\rho_j, L\}$$

where $K = [e^{-(\beta-\delta)} - L(1+q)]^{-1}$.

Let $\varepsilon > 0$. Choose $N = N(\varepsilon)$ so large that

$$(1-\chi)^{-1}\rho_n \leqslant \varepsilon/3K \quad \text{for} \quad n \geqslant N, \qquad \chi^{2N} q < \varepsilon/3, \qquad \chi^N < (NLK)^{-1}\varepsilon/3.$$

Then, we have for $n \geqslant 2N$

$$\lambda(n) \leqslant \chi^{2N} q + \chi^N K \sum_{j=0}^{N-1} \chi^{N-1-j} L + K \sum_{j=N}^{n} \chi^{n-1-j}\rho_j \leqslant \varepsilon/3 + \varepsilon/3 + \varepsilon/3 < \varepsilon. \tag{A.23}$$

Let now $n \leqslant t < n+1$. We have

$$\begin{aligned} \lambda(t) &\leqslant \frac{(|e^{A_3(t-n)}| + L)\,|y_3(n)| + \rho_n\,|y_{12}(n)|}{[|e^{-A_{12}(t-n)}|^{-1} - (1+q)L]\,|y_{12}(n)|}, \\ &\leqslant M_2\lambda(n) + M_1\rho(n), \end{aligned} \tag{A.24}$$

where $M_1 := [e^{-(\beta-\delta)L} - L(q+1)]^{-1}$ and $M_2 := (1+L)M_1$.

Let $\{\varepsilon_n\}$ be any sequence of positive reals satisfying $\varepsilon_n \to 0$ for $n \to \infty$. Define

$$r(t) := M_2\varepsilon_n + M_1\rho([t]) \qquad \text{for} \quad 2N(\varepsilon_n) \leqslant t < 2N(\varepsilon_{n+1}).$$

Then r depends on ρ only and satisfies (A.18); from (A.23) and (A.24) we obtain (A.20). ∎

A.4. Lemma

There exists a $q>0$, such that if η is sufficiently small, $x, x' \in \hat{\Gamma}(\eta)$, and $|x_3 - x_3'| \leqslant |x_{12} - x_{12}'|$, then

$$|P_3(\varphi_t(x) - x')| \leqslant q\,|(P_1 + P_2)(\varphi_t(x) - x')| \qquad (A.25)$$

whenever $t \geqslant 0$ and $\varphi_s(x) \in \Omega(\eta)$ for $0 < s \leqslant t$.

Proof

As usual we write $x(t) := \varphi_t(x)$ and assume that $\eta > 0$ is so small that Lemma A.1(i) holds true. We split the proof into three cases:

(a) $t \leqslant \tau$, $|x_{12} - x_{12}'| \leqslant \chi_1 \eta$,

(b) $t \leqslant \tau$, $|x_{12} - x_{12}'| > \chi_1 \eta$,

(c) $t > \tau$,

χ_1 and τ to be determined later. In each of the three cases we prove separately that for sufficiently small $\eta > 0$, a $q > 0$ satisfying the requirements of the lemma can be found.

Case (a)

We have

$$\frac{|x_3(t) - x_3'|^2}{|x_{12}(t) - x_{12}'|^2} \leqslant \frac{2[|x_3(t) - x_3|^2 + |x_3 - x_3'|^2]}{|x_{12}(t) - x_{12}|^2 + |x_{12} - x_{12}'|^2 - 2\langle x_{12}(t) - x_{12},\, x_{12} - x_{12}'\rangle}$$

We find $\chi_1 > 0$ and $0 < \tau \leqslant 1$, for which there exists a constant $\lambda < 1$ such that

$$\langle x_{12}(t) - x_{12},\, x_{12} - x_{12}'\rangle \leqslant \lambda\,|x_{12}(t) - x_{12}|\,|x_{12} - x_{12}'| \qquad (A.26)$$

provided $t \leqslant \tau$, $|x_{12} - x_{12}'| \leqslant \chi_1 \eta$ and $\eta > 0$ is sufficiently small.

Suppose for a moment that (A.26) holds. Since

$$|x_{12}(t) - x_{12}|\,|x_{12} - x_{12}'| \leqslant 1/2[|x_{12}(t) - x_{12}|^2 + |x_{12} - x_{12}'|^2],$$

we then have

$$\begin{aligned}\frac{|x_3(t) - x_3'|^2}{|x_{12}(t) - x_{12}'|^2} &\leqslant \frac{2[|x_3(t) - x_3|^2 + |x_3 - x_3'|^2]}{(1-\lambda)[|x_{12}(t) - x_{12}|^2 + |x_{12} - x_{12}'|^2]} \\ &\leqslant \frac{2}{(1-\lambda)}\left[\frac{|x_3(t) - x_3|^2}{|x_{12}(t) - x_{12}|^2} + \frac{|x_3 - x_3'|^2}{|x_{12} - x_{12}'|^2}\right] \\ &\leqslant \frac{2}{1-\lambda}\left[1 + \frac{|x_3(t) - x_3|^2}{|x_{12}(t) - x_{12}|^2}\right].\end{aligned}$$

For $t \leqslant 1$ we have by (2.6)

$$|x_3(t) - x_3| \leqslant |x(t) - x| \leqslant \int_0^t |Ax(x) + F(x(s))|\, ds \leqslant K_1 t\eta, \qquad \text{(A.27)}$$

where $K_1 := 2(|A| + L(\eta))$. Further, using the variation of constants formula we obtain

$$\begin{aligned} |x_{12}(t) - x_{12}| &\geqslant |x_2(t) - x_2| \geqslant |x_2| - |x_2(t)| \\ &\geqslant \eta - e^{-\gamma t}\eta - 2te^{|A_3|}\, L(\eta)\eta \geqslant K_2 t\eta \end{aligned}$$

with $K_2 = 1 - e^{-\gamma} - 2e^{|A_3|} L(\eta)$. Hence, if $\tau \leqslant 1$ and if $\eta > 0$ is so small that $e^{-\gamma} - 2e^{|A_3|} L(\eta) < 1$, for $0 \leqslant t \leqslant \tau$ we have

$$\frac{|x_3(t) - x_3'|^2}{|x_{12}(t) - x_{12}'|^2} \leqslant \frac{2}{1-\lambda}\left[1 + \frac{K_1}{K_2}\right],$$

i.e., (A.25) is satisfied with $q = (1-\lambda)^{-1}(1 + K_1^2/K_2^2)$.

To complete Case (a) it remains to find $1 \geqslant \tau > 0$ and $\chi_1 > 0$ such that (A.26) is satisfied for $0 \leqslant t \leqslant \tau$ and $|x_{12} - x_{12}'| \leqslant \chi_1 \eta$ for $\eta > 0$ sufficiently small.

Let π_η be the orthogonal projection of the neighborhood of the point x_{12} in $P_{12}\Gamma(\eta)$ into $T_{x_{12}} P_{12}\Gamma(\eta)$. Denote

$$f := \pi_\eta(x_{12} - x_{12}')\, |x_{12} - x_{12}'|^{-1}.$$

Then there exists a function $\omega\colon \mathbb{R}^+ \to \mathbb{R}^+$ (independent of η) such that $\omega(t) \to 0$ for $t \to 0$ and

$$|x_{12} - x_{12}' - f|x_{12} - x_{12}'|| \leqslant \omega(\eta^{-1}|x_{12} - x_{12}'|)\, |x_{12} - x_{12}'|.$$

We have

$$\begin{aligned} \langle x_{12}(t) - x_{12}, x_{12} - x_{12}' \rangle &= |x_{12} - x_{12}'|\, \langle x_{12}(t) - x_{12}, f \rangle \\ &\quad + \langle x_{12}(t) - x_{12}, x_{12} - x_{12}' - f\, |x_{12} - x_{12}'| \rangle, \end{aligned}$$

hence

$$\begin{aligned} &\langle x_{12}(t) - x_{12}, x_{12} - x_{12}' \rangle - |x_{12} - x_{12}'|\, \langle x_{12}(t) - x_{12}, f \rangle \\ &\quad \leqslant \omega(\eta^{-1}|x_{12} - x_{12}'|)\, |x_{12} - x_{12}'|\, |x_{12}(t) - x_{12}|. \end{aligned} \qquad \text{(A.28)}$$

We now estimate $\langle x_{12}(t) - x_{12}, f \rangle$. Since $\langle x_2, f \rangle = 0$, $\langle x_1, x_2 \rangle = 0$, we have

$$\begin{aligned} \langle x_{12}(t) - x_{12}, f \rangle &= |x_{12}(t) - x_{12}|^2 - \langle x_{12} - x_{12}(t), |x_2|^{-1} x_2 \rangle^2 \\ &= |x_{12}(t) - x_{12}|^2 - \eta^{-2} \langle x_2(t) - x_2, x_2 \rangle^2. \end{aligned} \qquad \text{(A.29)}$$

Further, for $0 \leqslant t \leqslant 1$, by (2.4), (2.6), and (A.27) we have

$$\begin{aligned}\langle x_2(t) - x_2, x_2 \rangle &= \left\langle \int_0^t [A_2 x_2(s) - F_2(x(s)], x_2 \right\rangle ds \\ &\geqslant t\langle A_2 x_2, x_2 \rangle - \eta \int_0^t [|A_2|\,|x_2(s) - x_2| + 2L(2\eta)]\, ds \\ &\geqslant t[\gamma - |A_2|\, tK_1 - L(\eta)]\eta^2.\end{aligned}$$

Hence, for $\tau \geqslant (1/4)|A_2|^{-1}K_1^{-1}\gamma$ and $L(\eta) < \gamma/4$ we have

$$\langle x_2(t) - x_2, x_2 \rangle \geqslant t(\gamma/2)\eta^2. \tag{A.30}$$

Substituting (A.30) into (A.29) we obtain

$$\langle x_{12}(t) - x_{12}, f \rangle^2 \leqslant |x_{12}(t) - x_{12}|^2 - t^2\gamma^2/4.$$

On the other hand, from (A.27) we obtain for $0 \leqslant t \leqslant 1$

$$t \geqslant (K_1\eta)^{-1}|x(t) - x| \geqslant (K_1\eta)^{-1}|x_{12}(t) - x_{12}|,$$

which implies

$$\langle x_{12}(t) - x_{12}, f \rangle^2 \leqslant \lambda_1 |x_{12}(t) - x_{12}|^2 \tag{A.31}$$

with $\lambda_1 := 1 - (1/4)K_1^{-2}\gamma^2 < 1$. From (A.28) and (A.31) it follows that

$$\langle x_{12}(t) - x_{12}, x_{12} - x'_{12} \rangle \leqslant \lambda |x_{12} - x'_{12}|\,|x_{12}(t) - x_{12}|$$

where $\lambda = \lambda_1 + \omega(\eta^{-1}|x_{12} - x'_{12}|)$. If χ_1 is chosen so small that $\omega(\chi_1) < 1 - \lambda_1$, we have $\lambda < 1$.

Case (b)

If $t < \tau$ and $|x_{12} - x'_{12}| \geqslant \chi_1\eta$ we have

$$\frac{|x_3(t) - x'_3|}{|x_{12}(t) - x'_{12}|} \leqslant \frac{2\eta}{\chi_1\eta - |x_{12}(t) - x_{12}|}.$$

Further, by (A.27) we have

$$|x_{12}(t) - x_{12}| \leqslant K_1\eta t.$$

If $\tau \leqslant (1/2)\chi_1 K_1^{-1}$, then

$$|x_3(t) - x'_3| \leqslant 4\chi_1^{-1}|x_{12}(t) - x'_{12}|,$$

i.e., (A.25) holds with $q := 4\chi_1^{-1}$.

Case (c)

Let $\tau \leqslant 1$ be given. For $t \geqslant 0$, $0 \leqslant s \leqslant \tau$ we obtain from (A.3) and the variation of constants formula

$$|x_2(t+s)| = \left| e^{A_2 s} x_2(t) + \int_0^s e^{A_2(s-\sigma)} F_2(x(t+\sigma))\, d\sigma \right|$$
$$\leqslant e^{-\gamma s} |x_2(t)| + 2se^{|A_2|} L(\eta)\eta. \tag{A.32}$$

In particular, for $t = \tau$ we have

$$|x_2(t)| \leqslant \chi_2 \eta \tag{A.33}$$

where $\chi_2 := e^{-\gamma\tau} + \tau e^{|A_2|} L(\eta) < 1$ for $\eta > 0$ sufficiently small.

Assume that (A.33) holds for $t = k\tau$. We prove that for $\eta > 0$ sufficiently small (A.33) extends to all $t \in [k\tau, (k+1)\tau]$.

From (A.32) and (A.33) applied to $t := k\tau$ we obtain for $0 \leqslant s \leqslant \tau$

$$x_2(k\tau + s) \leqslant e^{-\gamma s} \chi_2 \eta + se^{|A_2|} L(\eta)\eta \leqslant \chi_2 \eta,$$

provided $e^{-\gamma s}\chi_2 + se^{|A_2|} L(\eta) \leqslant \chi_2$ for $0 \leqslant s \leqslant \tau$, i.e., if $(1 - e^{-\gamma s})\chi_2 - se^{|A_2|} L(\eta) \geqslant 0$. Since $1 - e^{-\gamma s}$ is convex, this is true if $\eta > 0$ is chosen so small that $1 - e^{-\gamma\tau} \geqslant \tau e^{|A_2|} L(\eta)$.

By induction we obtain $|x_2(t)| \leqslant \chi_2 \eta$ for $t \geqslant \tau$. Hence, for $t \geqslant \tau$ we have

$$\frac{|x_3(t) - x_3'|}{|x_{12}(t) - x_{12}'|} \leqslant \frac{|x_3(t)| + |x_3'|}{|x_2'| - |x_2(t)|} \leqslant 2(1 - \chi_2)^{-1}.$$

Hence, (A.25) holds with $q := 2(1 - \chi_2)$. ∎

A.5. Lemma

For $\eta > 0$ sufficiently small we have

$$\lim_{t \to \infty} \Delta(t) = 0, \tag{A.34}$$

where

$$\Delta(t) := \sup \left\{ \frac{|P_3(S_{t+\tau}(x) - S_t(x'))|}{|(P_1 + P_2)(S_{t+\tau}(x) - S_t(x'))|} : \tau \geqslant 0;\ x, x' \in \hat{\Gamma}(\eta), \right.$$

$$\varphi_s(x) \in \Omega(\eta) \quad \text{for} \quad 0 < s \leqslant t + \tau, \qquad \varphi_\sigma(x') \in \Omega(\eta) \quad \text{for} \quad 0 < \sigma \leqslant t,$$

$$\left. |x_3 - x_3'| \leqslant |x_{12} - x_{12}'| \right\}. \tag{A.35}$$

Proof

First, we note that by Lemma A.4, for sufficiently small $\eta > 0$ there exists a $q > 0$ such that

$$|P_3(\varphi_\tau(x) - x')| \leqslant q\,|(P_1 + P_2)(\varphi_\tau(x) - x')|$$

provided $x, x' \in \hat{\Gamma}(\eta)$, and $|x_3 - x_3'| \leqslant |x_{12} - x_{12}'|$.

The assumptions of Lemma A.3 are satisfied for

$$y(t) := \varphi_{t+\tau}(x) - \varphi_t(x'),$$

with $L := L(\eta)$ and

$$\rho(t) = \sup\{|D_{x_{12}} R_3(s, (1-\vartheta)\,\varphi_{t+\tau}(x) + \vartheta\varphi_t(x'))| : t \geqslant 0, \quad 0 \leqslant s \leqslant 1,$$
$$0 \leqslant \vartheta \leqslant 1, \quad x, x' \in \operatorname{cl} \hat{\Omega}(\eta)\}.$$

To check that $\rho(t)$ satisfies (A.17), note that from (2.9) it follows that $D_{x_{12}} R_3(s, x_{12}, 0) = 0$ and, since R is C^1, $D_{x_{12}} R_3(s, x) \to 0$ for $x_3 \to 0$ uniformly for $0 \leqslant s \leqslant 1$ and $x \in \hat{\Omega}(\eta)$. Hence, (A.17) follows from Lemma A.1(iii). Now (A.34) is an immediate consequence of Lemma A.3. ∎

ACKNOWLEDGMENT

The author would like to express his gratitude to the anonymous referee for his careful reading of the manuscript and his helpful comments.

REFERENCES

Abraham, R., and Robbin, J. (1967). *Transversal Mappings and Flows*, Benjamin, New York.

Atkinson, F. V. (1964). *Discrete and Continuous Boundary Problems*, Academic Press, New York.

Brunovský, P. (1989). The maximal attractor of the scalar reaction diffusion equation. In Dafermos, C. M., Ladas, G., and Papanicolaou, G. (eds.), Differential Equations, Pure and Applied Mathematics 118, Marcel Dekker, New York, pp. 93–98.

Brunovský, P., and Chow, S. N. (1984). Generic properties of stationary state solutions of reaction diffusion equations. *J. Diff. Equat.* **53**, 1–23.

Brunovský, P., and Fiedler, B. (1986). Number of zeros on invariant manifolds in reaction diffusion equations. *Nonlin. Anal.* **10**: 179–193.

Brunovský, P., and Fiedler, B. (1988). Connecting orbits in scalar reaction diffusion equations. In Kirchgraber, U., and Walther, H. O. (eds.), *Dynamics Reported 1*, Wiley & Teubner, pp. 57–89.

Brunovský, P., and Fiedler, B. (1989). Connecting orbits in scalar reaction diffusion equations. II. The complete solution. *J. Diff. Equat.* **81**, 106–136.

Chow, S. N., and Lu, K. (1988). Invariant manifolds for flows in Banach spaces. *J. Diff. Equat.* **74**, 285–317.

Foias, C., Sell, G. R., and Temam, R. (1988). Inertial manifolds for nonlinear evolutionary equations. *J. Diff. Equat.* **73**, 309–353.

Fusco, G. (1987). Describing the flow on the attractor of one-dimensional reaction diffusion equations by systems of ODE. In Chow, S. N., and Hale, J. K. (eds.), *Dynamics of Infinite Dimensional Systems*, Computer and System Sciences 37, Springer, Berlin, pp. 113–122.

Hale, J. (1987). Some examples of infinite-dimensional systems. *Contemp. Math.* **58**(III), 173–182.

Hale, J., Magalhães, L., and Oliva, W. (1984). *An Introduction to Infinite Dimensional Dynamical Systems—Geometric Theory*, Appl. Math. Sci. 47, Springer, New York.

Hartman, P. (1964). *Ordinary Differential Equations*, Wiley, New York.

Henry, D. (1981). *Geometric Theory of Semilinear Parabolic Equations*, Lect. Notes Math. 840, Springer, New York.

Henry, D. (1985). Some infinite dimensional Morse-Smale systems defined by parabolic equations. *J. Diff. Equat.* **59**, 165–205.

Jolly, M. S. (1989). Explicit construction of an inertial manifold for a reaction diffusion equation. *J. Diff. Equat.* **78**, 220–261.

Kurzweil, J. (1970). Invariant manifolds I. *Comm. Math. Univ. Com.* **11**, 309–336.

Miklavčič, M. (1985). Stability for semilinear parabolic equations with noninvertible linear operator. *Pacif. J. Math.* **118**, 199–214.

Palis, J., and Melo, W. (1980). *Geometric Theory of Dynamical Systems*, Springer, New York.

P. Brunovský, P. Poláčik, B. Sandstede

Convergence in general periodic parabolic equations in one space dimension

Nonlinear Anal. 18(3) (1992), 209–215.

Nonlinear Analysis, Theory, Methods & Applications, Vol. 18, No. 3, pp. 209–215, 1992.
Printed in Great Britain.

0362-546X/92 $5.00+ .00

CONVERGENCE IN GENERAL PERIODIC PARABOLIC EQUATIONS IN ONE SPACE DIMENSION

P. BRUNOVSKÝ and P. POLÁČIK
Institute of Applied Mathematics, Comenius University, Mlynská Dolina, 84215 Bratislave, Czechoslovakia

and

B. SANDSTEDE
SFB 123, Universität Heidelberg, Im Neuenheimer Feld 294, 6900 Heidelberg, Germany

(*Received* 30 *July* 1990; *received for publication* 19 *April* 1991)

Key words and phrases: Periodic parabolic equation, convergence of solutions, periodic solutions.

IT HAS BEEN known for some time (see [8, 10, 14]) that scalar one-dimensional autonomous parabolic equations under separated boundary conditions have all bounded trajectories convergent. Recently, generalizations of this result to periodically timed-dependent equations have been established. Chen and Matano [5] have considered the equation

$$u_t = u_{xx} + f(t, u), \qquad t > 0, \qquad 0 < x < 1, \tag{1}$$

where f is of class C^2, $f(t + \tau, u) \equiv f(t, p)$ for some $\tau > 0$, under various types of boundary conditions (Dirichlet, Neumann, periodic). They have proved that any bounded solution of this boundary value problem converges to a τ-periodic solution of (1) and (2). In his thesis, Sandstede [13] extended this result by allowing f to depend on t, x, u and u_x. He proved the result for Dirichlet and, under some restrictions, for Neumann boundary conditions.

In this paper, we present a general convergence theorem with a simpler proof than the one in [13].

We consider a quasilinear parabolic equation

$$u_t = d(t, x, u, u_x)u_{xx} + f(t, x, u, u_x), \qquad t > 0, \qquad 0 < x < 1, \tag{2}$$

where $d, f \in C^2(\mathbb{R} \times [0, 1] \times \mathbb{R}^2, \mathbb{R})$, $d > 0$, are periodic in t with a common period $\tau > 0$. We consider either of the boundary conditions

$$u(t, i) = h_i(t), \qquad t > 0, \qquad i = 0, 1, \tag{3a}$$

$$u_x(t, i) = g_i(t, u(t, i)), \qquad t > 0, \qquad i = 0, 1. \tag{3b}$$

Here, $g_i(t, u)$ and $h_i(t)$, $i = 0, 1$, are C^2-functions, τ-periodic in t.

In the sequel the boundary conditions will be referred to as (3), assuming that only one of (3a), (3b) is chosen.

The problem (2), (3) is well posed on the Sobolev space $H^2 := H^2(0, 1)$ (see [1, 2]). For any $u_0(\cdot) \in H^2$ satisfying the compatibility conditions

$$u_0(i) = h_i(0), \qquad t > 0, \qquad i = 0, 1 \tag{4a}$$

or

$$u_{0x}(i) = g_i(0, u_0(i)), \qquad t > 0, \qquad i = 0, 1 \tag{4b}$$

[depending on whether we consider (3a) or (3b)], there exists a solution $u(t, \cdot)$ of (2), (3) with $u(0, \cdot) = u_0(\cdot)$. This solution is unique (up to the extension of the interval of existence) and depends continuously on u_0. Denoting the maximal interval of existence of $u(t, \cdot)$ by $[0, s_0)$, we have $s_0 = +\infty$ if $\|u(t, \cdot)\|_{H^2}$ stays bounded as $t \to s_0$. In the latter case, the set $\{u(t, \cdot): t > 0\}$ is relatively compact in H^2. By the regularity results of [1, 2], $u(t, x)$ is a classical solution (i.e. u_t, u_x, u_{xx} are continuous on $(0, s_0) \times [0, 1]$). Moreover, u_t has continuous derivative u_{tx}, hence (2) and the regularity of $d > 0$ and f imply that u_{xxx} is continuous.

The theorem we prove in this paper reads as follows.

THEOREM 1. Let $u(t, \cdot)$ be a bounded (in H^2) solution of (3), (4). Then there exists a τ-periodic solution $p(t, x)$ of (3), (4) such that

$$\lim_{t \to \infty} \|u(t, \cdot) - p(t, \cdot)\|_{H^2} = 0.$$

It will be useful to reformulate the conclusion of the theorem in terms of the Poincaré map T of the periodic problem (2), (3). By definition,

$$T(u_0) := u(\tau, \cdot)$$

if the solution $u(t, \cdot)$ with $u(0, \cdot) = u_0(\cdot)$ exists up to the time τ. The domain of definition of T is an open subset of the manifold

$$X := \{u_0 \in H^2: u_0 \text{ satisfies (4)}\}.$$

Clearly, $u(t, \cdot)$ is τ-periodic if and only if u_0 is a fixed point of T.

An obvious consequence of the conclusion of theorem 1 is that the sequence $T^n u_0 = u(n\tau, \cdot)$, $n = 0, 1, 2, \ldots$, converges to $p_0 \in X$. By the continuous dependence of the solutions of (2), (3) on initial conditions, the opposite is also true: if $T^n u_0 \to p_0(\cdot)$ [in which case $p_0(\cdot)$ is a fixed point of T], then the solution $p(t, x)$ of (2), (3) with $p(0, x) \equiv p_0(x)$ satisfies the conclusion.

The proof of the theorem is (of course) based on the properties of the zero number of solutions of a linearization of (2), (3). In fact, the quasilinearity of the equation is of no relevance. The arguments apply to any equation

$$u_t = F(t, x, u, u_x, u_{xx}), \qquad t > 0, \qquad 0 < x < 1, \tag{5}$$

with $F(t, x, u, p, q) \in C^2$, $F_q \in C^2$, F periodic in t, and $F_q \geq \beta > 0$ everywhere, provided the basic theory (existence, uniqueness, continuous dependence) and sufficient regularity (in general, continuity of u_{xxxx} is needed) is available. The reader interested in fully nonlinear equations is referred to [6, 7] and the references therein, where the basic properties are studied. Note that in the fully nonlinear case, compactness of the closure of a trajectory is not assured by its boundedness (so in formulations of convergence results, compactness must be assumed).

In order to prepare the proof of theorem 1, we now state a lemma which appears to be crucial.

Consider the linear equation

$$v_t = a(t, x)v_{xx} + b(t, x)v_x + c(t, x)v, \qquad t > 0, \qquad 0 < x < 1, \tag{6}$$

with the boundary conditions

$$v(t, i) = 0, \qquad i = 0, 1, \qquad t > 0, \tag{7a}$$

or

$$v_x(t, i) = \alpha_i(t)v(t, i), \qquad i = 0, 1, \qquad t > 0. \tag{7b}$$

LEMMA 1. Assume that a_t, a_x, a_{xx}, b_t, b_x and c are continuous on $[0, \infty) \times [0, 1]$, $a > 0$ everywhere, and that α_i, $i = 0, 1$, are bounded C^1 functions on $[0, \infty)$. Let $v(t, x) \not\equiv 0$ be a classical solution of (6), (7). Then there exists a t^* such that for any $t > t^*$ we have

$$v_x(t, 0) \neq 0, \text{ in the case of (8a) and}$$

$$v(t, 0) \neq 0, \text{ in the case of (8b).}$$

In the case of Dirichlet boundary conditions this lemma follows directly from the results of [3]. Indeed, by [3, theorem C], for $t > 0$, the "zero number"

$$z(v(t, \cdot)) := \sup\{k \in \mathbb{Z}: \text{there exist } 0 < x_1 < x_2 < \cdots < x_k < 1,$$
$$\text{such that } v(t, x_j)v(t, x_{j+1}) < 0, \text{ for } j = 1, 2, \ldots, k - 1\}$$

is finite and nonincreasing in t. Moreover, $z(v(t, \cdot))$ drops at any t such that $v(t, \cdot)$ has a multiple zero in $[0, 1]$. Since the integer value $z(v(t, \cdot)) \geq 0$ can drop only a finite number of times, there exists a t^* such that for $t > t^*$, $v(t, \cdot)$ has only simple zeros. In particular, (7a) implies that $u_x(t, 0) \neq 0$ for $t > t^*$.

In the case of the boundary condition (7b), lemma 1 is not so immediate. For the reader's convenience, its proof is included at the end of the paper.

The application of lemma 1 in the proof of theorem 1 is based upon the following observation: if u_1, u_2 are two solutions of (2), (3) then the difference $v := u_1 - u_2$ is a classical solution of (6), (7) with a, b and c defined by

$$a(t, x) = d(t, x, u_1, u_{1x}),$$

$$b(t, x) = \{f(t, x, u_1, u_{1x}) - f(t, x, u_1, u_{2x}) + (d(t, x, u_1, u_{1x}) - d(t, x, u_1, u_{2x}))u_{2xx}\}(u_{1x} - u_{2x})^{-1},$$

$$c(t, x) = \{f(t, x, u_1, u_{2x}) - f(t, x, u_2, u_{2x}) + (d(t, x, u_1, u_{2x}) - d(t, x, u_2, u_{2x}))u_{2xx}\}(u_1 - u_2)^{-1},$$

for $u_1 \neq u_2$, $u_{1x} \neq u_{2x}$, and extended continuously to the set where $u_1 = u_2$ or $u_{1x} = u_{2x}$. [For brevity we have omitted the argument (t, x).] In the case of Neumann boundary conditions [(3b) for u_1, u_2 and (7b) for v] we have

$$\alpha_i(t) = (g_i(t, u_1(t, i)) - g_i(t, u_2(t, i)))(u_1(t, i) - u_2(t, i))^{-1}.$$

By hypotheses and by the regularity properties of the solutions of (2), (3), the functions a, b, c and α_i satisfy the regularity assumptions of lemma 1. Moreover, α_1, α_2 are bounded if u_1, u_2 are.

We now prove theorem 1.

Let $u(t, x)$ be a bounded solution of (2), (3). As mentioned above, it suffices to prove that the sequence $T^n u(0, \cdot) = u(n\tau, \cdot)$, $n = 0, 1, \ldots$, is convergent. We first prove that the real sequence

$$\eta_n := (1 - \delta)u(n\tau, 0) + \delta u_x(n\tau, 0) \tag{8}$$

is convergent. Here $\delta = 1$ in the case of (3a) and $\delta = 0$ in the case of (3b).

Consider the function

$$v(t, x) := u(t + \tau, x) - u(t, x).$$

Due to periodicity, $u(t + \tau, x)$ satisfies (2), (3) [as does $u(t, x)$], hence $v(t, x)$ is a classical solution of some linear problem (6), (7). By lemma 1, unless $v \equiv 0$ (in which case u is τ-periodic and the assertion is trivial), there exists a t^* such that the function

$$t \mapsto (1 - \delta)v(t, 0) + \delta v_x(t, 0)$$

is of constant nonzero sign in (t^*, ∞).

Observe that

$$\eta_{n-1} - \eta_n = (1 - \delta)v(n\tau, 0) + \delta v_x(n\tau, 0).$$

Thus for $n > t^*\tau^{-1}$, η_n is a monotone sequence. Since η_n is bounded [because $u(t, \cdot)$ is bounded in H^2 and, thus, in C^1], it is convergent.

Denote

$$\eta_\infty := \lim_{n \to \infty} \eta_n. \tag{9}$$

We now prove that the ω-limit set $\omega(u)$, defined as the set of all accumulation points of $u(n\tau, \cdot)$ as $n \to \infty$, consists of a single point, i.e. $u(n\tau, \cdot)$ is convergent (recall that this sequence is relatively compact in the submanifold $X \subset H^2$). In the proof we use the obvious fact that

$$(1 - \delta)w(0) + \delta w_x(0) = \eta_\infty \tag{10}$$

for any $w(\cdot) \in \omega(u)$ [see (8), (9)].

Let $p_0(\cdot)$, $q_0(\cdot) \in \omega(u)$, and let $p(t, x)$, $q(t, x)$ be the solutions of (2), (3) with $p(0, x) \equiv p_0(x)$, $q(0, x) \equiv q_0(x)$. In order to prove that $p_0 = q_0$, we apply lemma 1 again, this time with the function

$$v(t, x) := p(t, x) - q(t, x).$$

By continuity of the Poincaré map T, we have $p(n\tau, \cdot) = T^n p_0 \in \omega(u)$ and, similarly, $q(n\tau, \cdot) \in \omega(u)$ for all n. Therefore, by (10),

$$(1 - \delta)p(n\tau, 0) + \delta p_x(n\tau, 0) \equiv \eta_\infty \equiv (1 - \delta)q(n\tau, 0) + \delta q_x(n\tau, 0).$$

Hence,

$$\delta v(n\tau, 0) + (1 - \delta)v_x(n\tau, 0) = 0 \qquad \text{for } n = 1, 2, \ldots.$$

By lemma 1, this is possible only if $v \equiv 0$. This shows that $p = q$ and completes the proof of the theorem.

We now prove lemma 1 for the boundary condition (7b). To this end we employ the following lemma which can be proved by adapting standard maximum principle arguments [4, 9, 11] in a straightforward way.

LEMMA 2. Let the functions a, b and c be defined on $D := [t_1, t_2] \times [x_1, x_2]$ with a_t, a_x, a_{xx}, b_t continuous and $a > 0$. Let $v(t, x)$ be a classical solution of (7) on D and for both $i = 0$ and $i = 1$ let one of the following conditions be satisfied:

(a) $u_x(t, x_i) \equiv 0$ for all $t \in [t_1, t_2]$,

(b) $u(t, x_i) \not\equiv 0$ for any $t \in [t_1, t_2]$.

Then, $z_{[x_1,x_2]}(u(t, \cdot))$ is a nonincreasing function of t.

Here $z_{[x_1,x_2]}$ stands for the "zero number" on the interval $[x_1, x_2]$ which is defined similarly to the zero number on $[0, 1]$.

Proof of lemma 1 *for* (8b). Let $v \not\equiv 0$ be a classical solution of (6), (7b) on $Q := (0, \infty) \times [0, 1]$. First we show that $z(v(t, \cdot)) < \infty$ for some $t_0 > 0$. This, in conjunction with the nonincrease of $z(v(t, \cdot))$ will imply that $z(v(t, \cdot))$ is constant at some interval (t^*, ∞). We then conclude the proof by showing that $v(t, 0) \not\equiv 0$ for $t > t^*$.

To see that $z(v(t, \cdot))$ is finite for $t > 0$, we use the following simple observation: arbitrarily near 0 there exists an open interval $U \subseteq (0, \infty)$ such that one of the following three alternatives holds:

(i) $v(t, i) \equiv 0$ for any $t \in U$, $i = 0, 1$;

(ii) $v(t, i) \not\equiv 0$ for any $t \in U$, $i = 0, 1$;

(iii) $v(t, 0) \equiv 0$ and $v(t, 1) \not\equiv 0$ for any $t \in U$;

(iv) $v(t, 0) \not\equiv 0$ and $v(t, 1) \equiv 0$ for any $t \in U$.

In the case of alternatives (i) and (ii), theorems C and D of [3] apply respectively to the solution $v(t, x)$ on $U \times [0, 1]$. By these theorems, $z(v(t, \cdot)) < \infty$ for $t \in U$. In the case of alternatives (iii) and (iv), the proofs of the above theorems have to be combined (cf. [3, pp. 81, 82]) to conclude the result.

In order to be able to apply lemma 2 in our next argument we "transform" the boundary conditions. To this end, consider the function

$$w(t, x) = \begin{cases} v(t, x)(1 + \xi(x)\alpha_0(t))^{-1} & \text{for } x \in [0, 1/2],\ t > 0, \\ v(t, x)(1 + \xi(x)\alpha_1(t))^{-1} & \text{for } x \in [1/2, 1],\ t > 0, \end{cases} \tag{11}$$

where $\xi(x)$ is a smooth function on $[0, 1]$ satisfying the following conditions:

$$\inf\{\xi(x)\alpha_i(t)\colon x \in [0, 1],\ t > 0,\ i \in \{0, 1\}\} > -1,$$

$$\xi \equiv 0 \text{ in a neighbourhood of } x = 1/2,$$

$$\xi(i) = 0,\ \xi_x(i) = 1 \text{ for } i := 0, 1.$$

It is easy to see that such a function ξ exists and that w is a classical solution of a linear equation

$$w_t = \tilde{a}w_{xx} + \tilde{b}w_x + \tilde{c}w,$$

where $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$ have the same regularity as a, b and c. Moreover, w satisfies

$$w_x(t, 0) = w_x(t, 1) = 0.$$

Thus, the transformation (11) leads to a boundary value problem (6), (7b) with

$$\alpha_1(t) \equiv 0, \qquad i := 0, 1. \tag{12}$$

Since v and w have the same zeros, this transformation shows that without loss of generality we may proceed in the proof for $v(t, x)$, assuming (12).

By lemma 2, $z(v(t, \cdot))$ is nonincreasing. Since $z(v(t, \cdot))$ is finite for $t > 0$, there is a t^* such that for $t > t^*$ we have

$$z(v(t, \cdot)) \equiv \text{const.}$$

We prove that $v(t, 0) \neq 0$ for $t > t^*$. Suppose the opposite holds, i.e. $v(t_1, 0) = 0$ for some $t_1 > t^*$. We show that this leads to a contradiction.

Since $z(v(t_1, \cdot)) < \infty$ and $v(t_1, \cdot) \not\equiv 0$ on $[0, 1]$ (otherwise $v(t, x) \equiv 0$ by the maximum principle), there exists an $x_0 \in (0, 1)$ such that

$$v(t_1, x_0) \neq 0 \qquad \text{and} \qquad v(t_1, x)v(t_1, x_0) \geq 0 \qquad \text{for } x \in [0, x_0]. \tag{13}$$

Assume, e.g. $v(t_1, x_0) > 0$ [the case $v(t_1, x_0) < 0$ is analogous]. Choose t_2, t_3 satisfying $t^* < t_2 < t_1 < t_3$ such that

$$v(t, x_0) > 0 \qquad \text{for } t \in [t_2, t_3].$$

By lemma 2, the functions

$$z_{[0,x_0]}(v(t, \cdot)), \qquad z_{[x_0,1]}(v(t, \cdot))$$

are both nonincreasing. Since we obviously have

$$z(v(t, \cdot)) = z_{[0,x_0]}(v(t, \cdot)) + z_{[x_0,1]}(v(t, \cdot)) \qquad \text{for } t \in (t_2, t_3)$$

and $z(v(t, \cdot))$ is constant for $t > t_2 > t^*$, $z_{[0,x_0]}(v(t, \cdot))$ must be constant as well. Hence, by (13),

$$z_{[0,x_0]}(v(t, \cdot)) \equiv 0 \qquad \text{for } t \in (t_2, t_3).$$

Consequently,

$$v(t, x) \geq 0 \qquad \text{on } Q_0 := [t_2, t_3] \times [0, x_0].$$

We see that 0 is the minimum of v in Q_0 and it is achieved at the boundary point (t_1, x_0). Now, the Neumann condition $u_x(t_1, x_0) = 0$ contradicts the Hopf boundary principle [12].

This contradiction shows that $v(t, 0) \neq 0$ for $t > t^*$ and lemma 1 is proved.

Acknowledgement—P. Brunovský and P. Poláčik were supported in part by the Institute for Mathematics and its Applications, University of Minnesota, with funds provided by the National Science Foundation.

REFERENCES

1. AMANN H., Quasilinear evolution equations and parabolic systems, *Trans. Am. math. Soc.* **293**, 191–227 (1986).
2. AMANN H., Dynamic theory of quasilinear parabolic equations, II: Reaction diffusion systems, *Diff. Integral Eqns* **3**, 13–75 (1990).
3. ANGENENT S., The zero set of a solution of a parabolic equation, *J. reine angew. Math.* **390**, 79–96 (1988).
4. BRUNOVSKÝ P. & FIEDLER B., Numbers of zeros on invariant manifolds in reaction–diffusion equations, *Nonlinear Analysis* **10**, 179–193 (1986).
5. CHEN X.-Y. & MATANO H., Convergence, asymptotic periodicity, and finite-point blow-up in one dimensional semilinear parabolic equations, *J. diff. Eqns* **78**, 160–190 (1989).
6. LUNARDI A., On the local dynamical system associated to a fully nonlinear abstract parabolic equations, in *Nonlinear Analysis and Applications* (Edited by V. LAKSHMIKANTHAM), pp. 319–326. Marcel Dekker, New York (1987).

7. LUNARDI A., Maximal space regularity in nonhomogeneous initial value parabolic problems, *Numer. Funct. Anal. Optimiz.* **10**, 323–349 (1989).
8. MATANO H., Convergence of solutions of one-dimensional semilinear parabolic equations, *J. Math. Kyoto Univ.* **18**, 221–227 (1978).
9. MATANO H., Nonincrease of the lap number of a solution for a one-dimensional semilinear parabolic equation, *J. Fac. Sci., Univ. Tokyo, Sec.* 1A **29**, 401–441 (1982).
10. MATANO H., Asymptotic behaviour of solutions of semilinear heat equations on S^1, in *Nonlinear Diffusion Equations and their Equilibrium States* (Edited by B. PELETIER and J. SERRIN), pp. 139–162. Springer, New York (1988).
11. NICKEL K., Gestaltaussagen über Lösungen parabolisher Differentialgleichungen, *J. reine angew. Math.* **211**, 78–94 (1962).
12. PROTTER M. H. and WEINBERGER H., *Maximum Principles in Differential Equations.* Prentice-Hall, Englewood Cliffs, NJ (1967).
13. SANDSTEDE B., ω-Limesmengen von Lösungen eindimensionaler Wärmeleitungsgleichungen mit periodischer Zeitabhängigkeit, Thesis, University of Heidelberg (1989).
14. ZELENYAK T. J., Stabilization of solutions of boundary value problems for a second order parabolic equation with one space variable, *J. diff. Eqns* **4**, 17–22 (1968).

P. Brunovský

Notes on chaos in the cell population partial differential equation

Nonlinear Anal. 7(2) (1983), 167–176.

Nonlinear Analysis, Theory, Methods & Applications, Vol. 7, No. 2, pp. 167–176, 1983.
Printed in Great Britain.

0362-546X/83/020167-10 $03.00/0

NOTES ON CHAOS IN THE CELL POPULATION PARTIAL DIFFERENTIAL EQUATION

PAVOL BRUNOVSKÝ
Institute of Applied Mathematics, Comenius University, 842 15 Bratislava, Czechoslovakia

(*Received in revised form* 15 *June* 1982)

Key words and phrases: First order partial differential equation, semiflow, chaos.

1. INTRODUCTION

IN [1], THE author investigates the differential equation

$$\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} = f(x, u), (t, x) \in D = [0, \infty) \times \Delta, \Delta = [0, 1]. \tag{1}$$

This equation describes the dynamics of growth of certain types of cell populations most prominent of which is the red blood cell population. It is shown in [1] that under certain natural conditions on c and f the equation (1) generates a semiflow $S_t, t \geqslant 0$ on $C_+(\Delta)$ (the space of nonnegative continuous functions on Δ) with an invariant set V_w on which the behaviour of the trajectories of S_t is chaotic in the sense of [2]. This means that S_t has a dense trajectory in V_w and each point of V_w is unstable (i.e. for each $v \in V_w$ there exists a neighbourhood U of $S_{(0,\infty)}v$ in $C(D)$ and a sequence $v_n \to v$ such that the trajectory of v_n leaves U for some $t \geqslant 0$).

The main purpose of this paper is to show that S_t exhibits also other features of chaos in V_w. Namely, there are periodic points of S_t of any basic period in V_w and the set of all periodic points of S_t is dense in V_w (Section 2).

For the proof a representation of S_t is employed which allows to prove the results on chaos of [1] in a more simple and transparent way. These proofs are presented in Section 3. Also, this technique helped to discover a small error in [1]. For the results on chaos of [1] to be true an additional (albeit also natural) assumption has to be added. We make this assumption in Sections 2 and 3. In Section 4 we discuss the modifications to be made if this additional assumption is dropped.

We keep all the notation of [1] in order to make it easier for the reader to relate the two papers. However, in order not to force the reader to look into [1] for every single concept or result we conclude this section by a list of assumptions and results of [1] used in the present paper.

Assumptions

A1. The functions c, f are continuously differentiable.
A2. $c(0) = 0$, $c(x) > 0$ for $x > 0$.
A3. There exists a $u_0 \in (0, 1]$ such that $f_u(0, u_0) < 0$, $f(0, u)(u - u_0) < 0$ for $u > 0$, $u \neq u_0$.
A4. $f(x, u) \leqslant k_1 u + k_2$ for some $k_1, k_2 \geqslant 0$ and all $x \in \Delta$, $u \geqslant 0$.
A5. $f(x, 0) = 0$ for all $x \in \Delta$.

Note that the assumptions **A1–A5** coincide with assumptions (16)–(18) in [1] with one difference:
A5 is somewhat sharper than the assumption
A5′. $f(x, 0) \geq 0$ for $x \in \Delta$ and $f(0, 0) = 0$
made in [1]. Also note that **A5** is satisfied if $f(x, u) = (p(x, u) - c(x))\, u$ as is the case if (1) models a reproductive, constantly differentiating cell population with proliferation rate p.

Results

Under the assumptions **A1–A4**, **A5′** the following results are proven in [1]:

R1. For $G \subset R^n$, $n > 0$, denote by $C_+(G)$, $C_+^1(G)$ the set of all nonnegative continuous and nonnegative continuously differentiable functions on G, respectively. For every $v \in C_+^1(\Delta)$, (1) has a unique solution u in $C_+^1(D)$ satisfying

$$u(x, 0) = v(x) \quad \text{for} \quad x \in \Delta. \tag{2}$$

A function $u \in C_+(D)$ is called generalized solution of (1) if it is a limit (uniform on compact subset of D) of solutions of (1). For each $v \in C_+(\Delta)$ there exists a unique generalized solution of (1) satisfying (2); henceforth we shall drop the adjective 'generalized'. The map $S: [0, \infty) \times C_+(\Delta) \to C_+(\Delta)$ defined by $S_t v(x) = u(t, x)$, where u satisfies (1), (2) is a continuous semiflow, i.e. $S_t: C_+(\Delta) \to C_+(\Delta)$ is continuous for each $t \geq 0$ and one has $S_0 = \mathrm{id.}$, $S_t \cdot S_s = S_{t+s}$ for each $t, s \geq 0$.

R2. Along the characteristics of (1) which are the curves $x = \varphi(t; t_0, x_0)$ satisfying the ordinary differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = c(x) \tag{3}$$

and the initial condition $x(t_0) = x_0$, the solution $u(t, x)$ of (1) satisfies the ordinary differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(\varphi(t; t_0, x_0), y) \tag{4}$$

with initial condition

$$y(0) = v(\varphi(0; t_0, x_0)); \tag{5}$$

the solution of (4), (5) is denoted by $\psi(t, \varphi(0; t_0, x_0), v(\varphi(0; t_0, x_0))$. This means that the solution u of (1) and (2) can be expressed by the formula

$$u(t, x) = \psi(t; \varphi(0; t, x), v(\varphi(0; t, x))). \tag{6}$$

For $\varphi(t; 0, x)$, write also $\varphi_x(t)$. It follows from **A2** that $\varphi_0(t) = 0$, $\varphi_x(t)$ is strictly increasing both in t and in x for $x > 0$, $\varphi_x^{-1}(1)$ is well defined continuous and decreasing for $0 < x \leq 1$.

R3. There exists a unique solution $w_0(x)$ of the stationary equation

$$c(x)\frac{\mathrm{d}u}{\mathrm{d}x} = f(x, u), \qquad x \in \Delta \tag{7}$$

satisfying $w_0(0) = u_0$. For each $v \in C_+(\Delta)$ such that $v(0) > 0$ one has $S_t v(x) \to w_0(x)$ for $t \to \infty$ uniformly in x.

R4. Let $V_0 = \{v \in C_+(\Delta): v(0) = 0\}$, $V_w = \{v \in V_0: v(x) < w_0(x) \text{ for } x \in \Delta\}$. The sets V_0, V_w are invariant for S_t and for each $v \in V_0$ there exists a $T_0 \geq 0$ such that $S_t v \in V_w$ for $t > T_0$.

We add two simple observations that will be used in the paper. Since $\varphi_0(t) = 0$ for all $t \geq 0$, it follows from (6) that a solution $u(t, x)$ of (1), (2) is well defined on $D^0 = [0, \infty) \times \Delta^0$ as soon as $v \in C_+(\Delta^0)$, where $\Delta^0 = (0, 1]$. In other words, the semiflow S_t can be extended to $C_+(\Delta^0)$; we denote this extended semiflow by S_t^0.

Further, since $u(t, x)$ is the solution of a first order ordinary differential equation along each characteristic, it follows from (6) that the semiflow S_t preserves ordering, i.e.

$$S_t v_1 \leqslant S_t v_2 \quad \text{for} \quad t \geq 0 \tag{8}$$

as long as $v_1 \leqslant v_2$, where $v_1 \leqslant v_2$ means $v_1(x) \leqslant v_2(x)$ for all $x \in \Delta$. This is true also for S_t^0.

2. EXISTENCE AND DENSITY OF PERIODIC POINTS

Throughout this and the following section assume **A1–A5**.

THEOREM 1. (a) For each $\tau \geq 0$ there is a continuum of periodic points of S_t in V_w of basic period τ. (b) The set of all periodic points of S_t is dense in V_w.

The basic tool of the proof of this theorem consists in the representation of S_t by the shift semigroup in $C_+[0, \infty)$. This representation is induced by the map $\Phi: C_+(\Delta) \to C_+[0, \infty)$ defined by

$$\Phi(v)(t) = (S_t v)(1).$$

Using (6) we can express Φ also by

$$\Phi(v)(t) = \psi\big(t; \varphi(0; t, 1), v(\varphi(0; t, 1))\big). \tag{9}$$

The family of shifts T_t, $t \geq 0$ defined by

$$(T_t g)(s) = g(t + s)$$

for $g \in C_+[0, \infty)$ is a semigroup and one has

$$T_t \Phi = \Phi S_t, \tag{10}$$

i.e. the diagram

$$\begin{array}{ccc} & \xrightarrow{S_t} & \\ \Phi \downarrow & & \downarrow \Phi \\ & \xrightarrow{T_t} & \end{array}$$

commutes.

Indeed,

$$\begin{aligned}(T_t \Phi(v))(s) &= \Phi(v)(s + t) = (S_{t+s} v)(1) = (S_s S_t v)(1) \\ &= \Phi(S_t v)(s).\end{aligned}$$

We can extend Φ to the map Φ_0 on $C_+(\Delta^0)$ by defining

$$\Phi_0(v)(t) = (S_t^0 v)(1).$$

Obviously, (10) holds with S_t, Φ replaced by S_t^0, Φ_0 respectively.

Let $g \in C_+[0, \infty)$. From (6) one immediately obtains $\Phi_0(v) = g$ if and only if

$$v(x) = \psi(-\varphi_x^{-1}(1); 1, g(\varphi_x^{-1}(1))) \quad \text{for} \quad x \in \Delta^0. \tag{11}$$

Using the argument leading to (8) one obtains from **A9** and (11)

$$v(x) \geqslant \psi(-\varphi_x^{-1}(1); 1, 0) = 0.$$

Thus we have

LEMMA 2.1. The map $\Phi_0 : C_+(\Delta^0) \to C_+[0, \infty)$ has an inverse which can be expressed by the formula (11).

Note that $\Phi_0^{-1}(g)$ is not necessarily in $C_+(\Delta)$ for an arbitrary $g \in C_+[0, \infty)$ since $\Phi_0^{-1}(g)$ may not have a limit for $x \to 0$.

As a consequence of **R3** one obtains immediately

LEMMA 2.2. Let $v \in C_+(\Delta)$ satisfy $v(0) > 0$. Then, $\Phi(v)(t) \to w_0(1)$ for $t \to 0$.

LEMMA 2.3. Let $g \in C_+[0, \infty)$ and let

$$g(t) \leqslant w_0(1) - \eta \tag{12}$$

for some $\eta > 0$ and each $t \geqslant 0$. Then $g \in \Phi(V_w)$.

Proof. Obviously, it suffices to prove

$$\lim_{x \to 0} \Phi_0^{-1}(g)(x) = 0 \tag{13}$$

since then $g = \Phi(v)$, where

$$v(x) = \begin{cases} \Phi_0^{-1}(g)(x) & \text{for } x \in \Delta^0 \\ 0 & \text{for } x = 0 \end{cases}$$

is from V_w. ■

To prove (13) we first introduce the following notation which will be used throughout the paper:

For any $c \geqslant 0$ we denote by $\mathbf{c}$ the constant function on Δ with value c and $h_c(t) = \Phi(\mathbf{c})(t)$.

Let now $\varepsilon > 0$. Since by lemma 2.2. $\lim_{t \to \infty} h_\varepsilon(t) = w_0(1)$, there exists a $t_0 > 0$ such that for $t > t_0$ one has

$$h_\varepsilon(t) > w_0(1) - \eta \geqslant g(t).$$

Let $x_0 = \varphi(0; t_0, 1)$. For $x < x_0$ one has $\varphi_x^{-1}(1) > t_0$, and, consequently, by (11).

$$\Phi_0^{-1}(g)(x) = \psi(-\varphi_x^{-1}(1); 1, g(\varphi_x^{-1}(1))) < \Phi_0^{-1}(h_\varepsilon(\varphi_x^{-1}(1))) = \varepsilon$$

Since $\varepsilon > 0$ was arbitrary this proves (13).

Since for $g \in C_+[0, \infty)$ periodic with values in $[0, w_0(1))$ there is always an $\varepsilon > 0$ such that (12) holds we have

COROLLARY 2.1. The function $g \in C_+[0, \infty)$ with values in $[0, w_0(1))$ is periodic with prime period $\tau \geq 0$ if and only if $\Phi^{-1}(g)$ is a periodic point of S_t in V_w with basic period τ. In particular, all the solutions of the stationary equation (7) in V_w are obtained as pre-images of constant functions $< w_0(1)$ under Φ.

LEMMA 2.4. For each $0 < \varepsilon < \inf_{0 \leq x \leq 1} w_0(x)$ there exists a $\tau_\varepsilon > 0$ such that $h_\varepsilon(s + t) \leq h_\varepsilon(s)$ for each $s \geq 0$, $t \geq \tau_\varepsilon$.

Proof. By **R3**, there exists a $\tau_\varepsilon > 0$ such that $S_t\varepsilon > \varepsilon$ for all $t \geq \tau_\varepsilon$. Hence, for $t \geq \tau_\varepsilon$ we have

$$h_\varepsilon(s + t) = (T_t h_\varepsilon)(s) = \Phi(S_t\varepsilon)(s) \geq \Phi(\varepsilon)(s) = h_\varepsilon(s). \quad \blacksquare$$

Proof of theorem 2.1. Part (a) is an immediate consequence of corollary 2.1.

To prove (b) take any function v in V_w and choose an $\varepsilon > 0$. Denote $g = \Phi(v)$. Let $\delta > 0$ be such that $v(x) < \varepsilon$ for $x < \delta$, so

$$g(t) < h_\varepsilon(t) \quad \text{for} \quad t \geq t_1 = \varphi_\delta^{-1}(1). \tag{14}$$

Let $t_2 > \max\{t_1, \tau_\varepsilon\}$ be such that

$$h_\varepsilon(t) \geq \max_{0 \leq t \leq t_1} g(t), \tag{15}$$

for $t \geq t_2$, τ_ε being as in lemma 2.4.

From (14), (15) it follows that there exists a continuous function $\tilde{g} \in C_+[0, t_2]$ such that

$$\tilde{g}(t) = g(t) \quad \text{for} \quad 0 \leq t \leq t_1, \tag{16}$$

$$\tilde{g}(t) < h_\varepsilon(t) \quad \text{for} \quad t_1 \leq t \leq t_2, \tag{17}$$

$$\tilde{g}(t_2) = g(0).$$

Define $k \in C_+[0, \infty)$ by

$$k(t) = \tilde{g}(t - nt_2) \quad \text{for} \quad t \in [nt_2, (n + 1)t_2].$$

Then, k is periodic with period t_2 and, by lemma 2.3., there is a $z \in V_w$ such that $k = \Phi(z)$. From (14) and (15) we obtain

$$z(x) = v(x) \quad \text{for} \quad \delta \leq x \leq 1, \tag{18}$$

$$|z(x)| < \varepsilon \quad \text{for} \quad \varphi(0; t_2, 1) \leq x \leq \delta. \tag{19}$$

Let $n \geq 1$. For $nt_2 + t_1 \leq t \leq (n + 1)t_2$ we obtain by lemma 2.4 and (14)

$$k(t) = \tilde{g}(t - nt_2) \leq h_\varepsilon(t - nt_2) \leq h_\varepsilon(t); \tag{20}$$

for $nt_2 \leqslant t \leqslant (n+1)t_2$, (20) follows immediately from (15). Consequently, (19) extends to all $0 \leqslant x \leqslant \delta$ and we have

$$|z(x) - v(x)| \leqslant |z(x)| + |v(x)| \leqslant 2\varepsilon$$

for $0 \leqslant x \leqslant \delta$. This, together with (18), proves (b). ■

3. EXISTENCE OF A DENSE TRAJECTORY AND INSTABILITY

Using the representation of S_t by T_t developed in Section 2 we now present an alternative proof of theorem 3 of [1]. That is, we prove

(a) every point $v \in V_w$ is unstable;

(b) there exists a $v \in V_w$ such that the orbit of v is dense in V_w.

Proof of (a). Let $v \in V_w$, $g = \Phi(v)$, $0 < a < w_0(1)$. Choose an $\varepsilon < 0$.

Let $\delta > 0$ be such that $v(x) < \varepsilon$ for $x \leqslant \delta$. Let $t_1 \geqslant \varphi_\delta^{-1}(1)$ be such that

$$h_\varepsilon(t) > a \tag{21}$$

for $t \geqslant t_1$.

We now construct a function $k \in C_+[0, \infty)$ as follows: We define

$$k(t) = g(t) \quad \text{for} \quad 0 \leqslant t \leqslant t_1$$

$$k(t_1 + j) = \begin{cases} a & \text{if} \quad g(t_1 + j) < \dfrac{a}{2} \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, 2, 3, \ldots$$

and we extend k to the interior of the intervals between the points $t_1 + j$ in such a way that k will be nonnegative continuous and its graph will lie below the graph of h_ε for $t_1 \leqslant t \leqslant t_1 + 1$ and below a for $t > t_1 + 1$. Then, we have

$$k(t) \leqslant h_\varepsilon(t) \quad \text{for} \quad t \geqslant t_1 \tag{22}$$

and

$$|k(t_1 + j) - g(t_1 + j)| \geqslant \frac{a}{2} \quad \text{for} \quad j = 1, 2, \ldots \tag{23}$$

By lemma 2.3., there exists a $z \in V_w$ such that $k = \Phi(v)$. Now, (23) can be rewritten as

$$|(S_{t_1+j}v)(1) - (S_{t_1+j}z)(1)| \geqslant \frac{a}{2}. \tag{24}$$

Also, we have from (18), (19)

$$z(x) = v(x) \quad \text{for} \quad \varphi(0; t_1, 1) \leqslant x < 1 \tag{25}$$

while

$$|z(x) - v(x)| \leqslant |z(x)| + |v(x)| \leqslant \varepsilon + \varepsilon = 2\varepsilon$$

for $0 \leqslant x < \varphi(0; t_1, 1)$. Since $\varepsilon > 0$ was arbitrary, (24)–(26) proves (a). ■

Proof of (b). Let $\{v_n\}_{n=1}^{\infty}$ be a dense subset in V_w and let $\varepsilon_n \searrow 0$ for $n \to \infty$. Denote $g_n = \Phi(v_n)$. By lemma 3.2., there exists a sequence $\{t_n\}$ such that

$$t_1 = 0, \; t_{n+1} \geq t_n + 1. \tag{26}$$

$$h_{\varepsilon_j}(t_{n+1} - t_n) \geq \varepsilon_{j+1} (= h_{\varepsilon_{j+1}}(0)) \quad \text{for} \quad 0 \leq j \leq n \tag{27}$$

$$g_n(t) \leq h_{\varepsilon_j}(t + t_n - t_j) \quad \text{for all} \quad t \quad \text{and all} \quad 1 \leq j < n \tag{28}$$

$$g_n(t) \leq h_{\varepsilon_n}(t + t_n) \quad \text{for all} \quad t \geq 0. \tag{29}$$

First we note that a sequence of continuous functions $\tilde{g}_n \in C_+[0, t_{n+1} - t_n]$ can be found such that $\tilde{g}_n(t) = g_n(t)$ for $0 \leq t \leq t_{n+1} - t_n - 1$, $\tilde{g}_n(t_{n+1} - t_n) = g_{n+1}(0)$ and the inequalities (27)–(29) remain valid with g_n replaced by $\tilde{g}_n$ and t restricted to $t_{n+1} - t_n$ (we shall refer to them as $(\widetilde{27})$–$(\widetilde{29})$, respectively). We define

$$k(t) = \tilde{g}_n(t - t_n) \quad \text{for} \quad t_n \leq t \leq t_{n+1}.$$

Obviously, $k \in C_+[0, \infty)$ and $k(t) < w_0(1)$ for $0 \leq t < \infty$. Further, we have by $(\widetilde{29})$

$$k(t) \leq h_{\varepsilon_n}(t) \quad \text{for } t_n \leq t \leq t_{n+1}$$

and, by (11),

$$\Phi_0^{-1}(k)(x) \leq \varepsilon_n \quad \text{for} \quad \varphi(0, t_{n+1}, 1) \leq x \leq \varphi(0, t_n, 1).$$

Consequently, $\lim_{x \to 0} \Phi_0^{-1}(k)(x) = 0$ and $k \in \Phi(z)$ for some $z \in V_w$.

Now, we have

$$(T_{t_n}k)(t) = g_n(t) \quad \text{for} \quad 0 \leq t \leq t_{n+1} - t_n - 1. \tag{30}$$

The inequalities $(\widetilde{27})$ and $(\widetilde{28})$ can be transcribed into

$$(T_{t_n}k)(t) \leq h_{\varepsilon_n}(t) \quad \text{for} \quad t \geq t_{n+1} - t_n - 1 \tag{31}$$

($(\widetilde{27})$ yields (31) for $t_{n+1} - t_n - 1 \leq t \leq t_{n+1} - t_n$ while $(\widetilde{28})$ yields (31) for $t \geq t_{n+1}$). From (30) and (31) we have

$$(S_{t_n}z)(x) = v_n(x) \quad \text{for} \quad (0; t_{n+1} - t_n - 1, 1) \leq x \leq 1) \tag{32}$$

$$(S_{t_n}z)(x) \leq \varepsilon_n \quad \text{for} \quad 0 \leq x \leq \varphi(0; t_{n+1} - t_n - 1, 1). \tag{33}$$

Also, from (27) we have

$$v_n(x) \leq \varepsilon_n \quad \text{for} \quad 0 \leq x \leq \varphi(0; t_{n+1} - t_n - 1, 1). \tag{34}$$

From (32)–(34) it follows

$$|(S_{t_n}z)(x) - v_n(x)| \leq 2\varepsilon_n \quad \text{for all} \quad x \in \Delta$$

which completes the proof. ■

Remark. It is easy to see that the function z giving the initial point of the dense trajectory in V_w can be constructed to be C^1 hence yielding a continuously differentiable solution of (1). This is true also for the functions z_n in part (a) and the periodic points of part (b) of theorem 2.1.

4. THE CASE $f(x, 0) \not\equiv 0$

Throughout this section we assume **A1–A4**, **A5′**. First we show that if **A5** is not satisfied there cannot be chaos in all of V_w.

PROPOSITION 4.1. Let $f(x_0, 0) > 0$ for some $x_0 \in \Delta$. Then V_w does not admit a dense trajectory.

LEMMA 4.1. For each $0 \leqslant t_1 \leqslant t_2$ one has

$$0 \leqslant S_{t_1}\mathbf{0} \leqslant S_{t_2}\mathbf{0}. \tag{35}$$

Proof. From (6) it follows

$$(S_t\mathbf{0})(x) \geqslant 0 \quad \text{for} \quad (t, x) \in D. \tag{36}$$

From (8), (36) and the semigroup property of S_t it follows

$$(S_{t_2}\mathbf{0})(x) = (S_{t_1}S_{t_2-t_1}\mathbf{0})(x) \geqslant (S_{t_1}\mathbf{0})(x). \quad \blacksquare$$

COROLLARY 4.1. Under the condition of proposition 4.1 there is a neighbourhood U of x_0 in Δ such that

$$(S_t\mathbf{0})(x) > 0 \quad \text{for each} \quad x \in U \quad \text{and} \quad t > 0. \tag{37}$$

Proof of proposition 4.1. Choose any $\tau > 0$ and denote $z = \frac{1}{2}S_\tau\mathbf{0}$. By corollary 4.1 we have $z \neq 0$. Assume $v \in V_w$ has a dense trajectory in V_w. Since $v \geqslant 0$, by (8) and lemma 4.1 we have

$$S_t v \geqslant 2z \quad \text{for all} \quad t \geqslant \tau. \tag{38}$$

Let $Z = \{w \in C_+(\Delta)\colon w(x) \leqslant z(x) \text{ for } x \in \Delta\}$. Since $z \neq 0$, $Z \neq \emptyset$. By (38), we have for all $\zeta \in Z$ and $t \geqslant \tau$

$$\sup_{x\in\Delta} |(S_t v)(x) - \zeta(x)| \geqslant \sup_{x\in\Delta} |2z(x) - z(x)| > 0.$$

Thus, in order that $S_{[0,\infty)}v$ be dense in V_w, $S_{[0,\tau]}v$ must be dense in Z. This, however, is easily seen to be impossible since $S_{[0,\tau]}v$ is compact in $C(\Delta)$ and does not contain all of Z. The compactness of $S_{[0,\tau]}v$ follows e.g. from the expression (6) from which one immediately concludes that the family of functions $\{S_t v\colon 0 \leqslant t \leqslant \tau\}$ is closed, uniformly bounded and equicontinuous.

Proposition 4.1 decides the question whether **A5** is necessary for the results on chaos to hold in their original form. Still, the results of [1] and Section 2 on chaos remain valid under **A5′** with V_w replaced by its invariant subset which we denote by W. To define W we need.

PROPOSITION 4.2. There exists a pointwise limit

$$w_1(x) = \lim_{t\to\infty} (S_t\mathbf{0})(x).$$

The function w_1 is a solution of the stationary equation (7) on Δ^0 satisfying

$$0 \leqslant w_1(x) \leqslant w_0(x) \quad \text{for} \quad x \in \Delta^0. \tag{39}$$

Proof. The existence of a pointwise limit w_1 of $S_t\mathbf{0}$ for $t \to \infty$ satisfying (39) is an immediate

consequence of lemma 5. It remains to prove that w_1 is a solution of (7) on Δ^0. For the idea of this proof the author is indebted to J. Kačur.

Denote $u(t, x) = (S_t\mathbf{0})(x)$. For this proof we write (1), (7) in the form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(c(x)u) = q(x, u), \tag{40}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(c(x)u) = q(x, u), \tag{41}$$

respectively, with $q(x, u) = f(x, u) + c'(x)u$.

Let $t \in [0, \infty)$, $x \in (0, 1]$. By integrating (40) we obtain

$$\int_1^x [u(t+1, \xi) - u(t, \xi)]\,\mathrm{d}\xi + \int_t^{t+1} [c(x)u(\sigma, x) - c(1)u(\sigma, 1)]\,\mathrm{d}\sigma$$

$$= \int_t^{t+1}\int_1^x q(\xi, u(\sigma, \xi))\mathrm{d}\xi\,\mathrm{d}\sigma. \tag{42}$$

Since $0 \leqslant u(t, x) \leqslant w_0(x)$ for all $(t, x) \in D$, by Lebesgue's convergence theorem we can pass to the limit for $t \to \infty$ in (42) to obtain

$$\int_t^{t+1} [c(x)w_1(x) - c(1)w_1(1)]\mathrm{d}t = \int_t^{t+1}\int_1^x q(\xi, w_1(\xi))\mathrm{d}\xi\,\mathrm{d}\sigma$$

and, consequently,

$$c(x)w_1(x) - c(1)w_1(1) = \int_1^x q(\xi, w_1(\xi))\mathrm{d}\xi. \tag{43}$$

From (43) it follows that w_1 is absolutely continuous on Δ^0. Thus, we can differentiate (43) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}(c(x)w_1(x)) = q(x, w_1(x)) \tag{44}$$

which completes the proof. ■

Now, denote

$$W = \{v \in V_w\colon v(x) \geqslant w_1(x) \quad \text{for} \quad x \in \Delta\}. \tag{45}$$

One sees immediately that W is invariant. It is also attractive in V_0 but, unlike V_w, only in a 'pointwise' sense: the graphs of the upper and lower pointwise limits of $S_t v$ for $t \to \infty$ lie between the graphs of w_1 and w_0, for each $v \in V_0$. This follows immediately from **R4** and

$$\liminf_{t\to\infty} (S_t v)(x) \geqslant \liminf_{t\to\infty} (S_t\mathbf{0})(x) = w_1(x) \quad \text{for} \quad x \in \Delta.$$

The map Φ maps w_1 into the constant $w_1(1)$. If one replaces $C_+[0, \infty)$ by its subset of functions with values $\geqslant w_1(1)$, lemma 2.1 obviously holds true and one can repeat the arguments of Sections 2 and 3 almost literally to obtain.

THEOREM 4.1. The set W defined by (45) is invariant and pointwise attractive in V_0. Also, S_t is chaotic in W in the sense of theorem 3 of [1] and theorem 2.1.

It should be noted that the chaotic set W may very well be empty. Obviously, W is non-empty if and only if (7) has a non-negative solution w_2 on Δ satisfying $w_2(0) = 0$. Indeed, every non-negative solution of (7) on Δ^0 majorized by w_0 and different from w_0 vanishes at 0 (lemma 2.3.); if w_2 exists one has $w_1(x) \leqslant w_2(x) < w_0(x)$ for $x \in \Delta$. It follows that the question, whether W is empty or not, is decided by the local behaviour of f and c at $(0, 0)$.

For example, W is non-empty if $f(x, 0)$ vanishes in some right neighbourhood of 0. On the other hand, take $f(x, u) = x^2 + u^2$, $c(x) = x^2$ for $x \geqslant 0$, $u \geqslant 0$ small. All integral curves of the equation

$$x^2 \frac{\mathrm{d}u}{\mathrm{d}x} = u^2 + x^2 \tag{46}$$

passing through points (x, u) with $x > 0$, $u \geqslant 0$ are given in parametric form by

$$x(s) = \mathrm{d} \exp[2.3^{-1/2} \arctan(3^{-1/2}(2s - 1))]$$
$$u(s) = sx(s) \quad (-\infty < s < \infty)$$

with $d > 0$. It can be readily seen that none of these curves approaches the point $(0, 0)$, so (46) has no solution with $u(0) = 0$. Consequently, W is empty for any extensions of f, c satisfying **A1–A4**, **A5′**.

REFERENCES

1. LASOTA A., Stable and chaotic solutions of a first-order partial differential equation, *Nonlinear Analysis* **5**, 1181–1193 (1981).
2. AUSLANDER J. & YORKE J., Interval maps, factor of maps and chaos, *Tohoku math. J.* Ser. II, **32**, 177–188 (1980).

A. Brunovská, M. Morbidelli, P. Brunovský

Optimal catalyst pellet activity distributions for deactivating systems

Chemical Engineering Science 45(4) (1990), 917–925.

Chemical Engineering Science, Vol. 45, No. 4, pp. 917–925, 1990.
Printed in Great Britain.

0009-2509/90 $3.00 + 0.00

OPTIMAL CATALYST PELLET ACTIVITY DISTRIBUTIONS FOR DEACTIVATING SYSTEMS

ALENA BRUNOVSKÁ†
Department of Organic Technology, Slovak Institute of Technology, Radlinského 9, 812 37 Bratislava, Czechoslovakia

MASSIMO MORBIDELLI
Department of Chemical Engineering and Materials Science, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy

and

PAVOL BRUNOVSKÝ
Institute of Applied Mathematics, Comenius University, Mlynská dolina, 842 15 Bratislava, Czechoslovakia

(*Received* 23 *January* 1989; *accepted* 13 *July* 1989)

Abstract—The optimal activity distribution in catalyst pellets for reacting systems which undergo deactivation is analysed. A general optimality criterion is developed, which allows one to conclude that, under quite general conditions, the optimal activity distribution is of the Dirac-delta type.

INTRODUCTION

As an important component of catalyst design, optimal active catalyst distribution in the porous structure of the inert support has received considerable attention in the literature. The main results have been recently reviewed by Dougherty and Verykios (1987). Most previous works have been devoted to the problem of increasing the effectiveness factor or selectivity in some specific reacting systems. Only a few papers dealt with catalytic systems undergoing deactivation, again considering specific cases (De Lancey, 1973; Corbett and Luss, 1974; Becker and Wei, 1977a, b). For example, De Lancey (1973) estimated the optimal activity distribution for an isothermal first-order reaction and homogeneous poisoning using Pontrjagin's maximum principle.

In this paper the problem of catalyst design for systems which undergo deactivation is analysed. In particular, we refer to noble-metal catalysts dispersed within a particle of inert support. These systems are widely used in industry for hydrogenation and oxidation reactions, which constitute intermediate steps in the production of a variety of chemical products. In most cases the catalyst undergoes deactivation. Its replacement or, when possible, its regeneration, constitutes a significant part of the production cost.

The aim of the present paper is to optimize the catalyst pellet performance by suitably locating the active element within the particle support. To this aim, a general optimality condition is derived which allows to solve the optimization problem under quite general conditions including any rate expression for both the main reaction and the poisoning process under non-isothermal conditions.

THE OPTIMIZATION PROBLEM

The catalyst which is progressively poisoned with operating time has to be periodically replaced or regenerated, depending upon whether the poisoning is irreversible or reversible. The duration of the operating time and the values of the effectiveness factor as a function of time depend upon the active-catalyst distribution within the support. In general, by locating the active catalyst inside the pellet it is possible to increase the resistance against deactivation, i.e. to increase the duration of the operating time. On the other hand, at least for positive-order reactions, the maximum value of the effectiveness factor is obtained when the active catalyst is located at the external surface (Morbidelli *et al.*, 1985; Chemburkar *et al.*, 1987). Thus, an economic criterion is needed to define the optimal active catalyst distribution. A reasonable one is profit per time:

$$\text{profit per time} = \frac{\text{price of the product} - \text{cost of the catalyst}}{\text{operating time}} = \frac{\alpha_1 \int_0^{\tau^*} \eta \, d\tau - \alpha_2}{\tau^*} \quad (1)$$

where α_1 and α_2 are weighting coefficients proportional to the price of the product and to the cost of the catalyst, respectively, τ^* is the operating time, and η is the effectiveness factor.

The aim of this work is to determine the initial pellet activity distribution $a(\varphi, 0)$ and the operating time τ^* for which the maximum value of the following objective function, proportional to the profit per time,

†Author to whom correspondence should be addressed.

defined above:

$$\mathcal{J}[a(\varphi, 0), \tau^*] = \frac{\gamma \int_0^{\tau^*} \eta \, d\tau - 1}{\tau^*} \tag{2}$$

($\gamma = \alpha_1/\alpha_2$) is obtained. We optimize over the class of all possible distributions of the same amount of active catalyst. Note that we admit also distributions concentrating the active catalyst into isolated points which are represented by Dirac-delta functions.

THE BASIC EQUATIONS

Let us consider a catalyst pellet in which an irreversible reaction is taking place together with irreversible adsorption of catalyst poison. Since the rate of the poison adsorption is usually considerably lower than that of the catalytic reaction (the form of which may otherwise be arbitrary) the quasi-steady-state approximation can be safely adopted. In addition, we assume negligible external resistances to mass and heat transport. The catalyst activity distribution is a function of location and time, and is defined as the ratio between the local concentration of available catalytically active sites and its volume-averaged initial value:

$$a(\varphi, \tau) = \sigma(\varphi, \tau)/\bar{\sigma} \tag{3}$$

where

$$\bar{\sigma} = (n+1) \int_0^1 \sigma(\varphi, 0) \varphi^n \, d\varphi. \tag{4}$$

Under these conditions, the model equations in dimensionless form are as follows:

Mass balance of the reactant

$$\nabla^2 Y = \Phi^2 R \tag{5}$$

Mass balance of the poison

$$\nabla^2 Y_p = \Phi_p^2 R_p \tag{6}$$

Energy balance

$$\nabla^2 \upsilon = -\beta \Phi^2 R \tag{7}$$

with boundary conditions

$$\varphi = 0: \; \partial Y/\partial \varphi = \partial Y_p/\partial \varphi = \partial \upsilon/\partial \varphi = 0$$

$$\varphi = 1: \; Y = Y_p = \upsilon = 1. \tag{8}$$

The deactivation reaction is accounted for by a balance of the active sites, which in terms of the activity distribution function reduces to

$$\frac{\partial a}{\partial \tau} = -R_p \tag{9}$$

with initial condition

$$a = a(\varphi, 0) \quad \text{at } \tau = 0 \tag{10}$$

where the initial activity distribution has to satisfy the following constraint arising from its definition (3) and eq. (4):

$$(n+1) \int_0^1 a(\varphi, 0) \varphi^n \, d\varphi = 1. \tag{11}$$

Note that no restrictions are imposed on the expressions of the rates of the reaction and the poisoning processes, for which the following general form is assumed:

$$R = R(Y, Y_p, a, \upsilon); \; R_p = R_p(Y, Y_p, a, \upsilon). \tag{12}$$

The effectiveness factor η is normalized with respect to the initial value of the reaction rate computed at surface conditions and to the initial activity distribution:

$$\eta = \int_0^1 \varphi^n R \, d\varphi \Big/ \int_0^1 a(\varphi, 0) \varphi^n \, d\varphi$$

$$= (n+1) \int_0^1 \varphi^n R \, d\varphi = \bar{R} \tag{13}$$

thus representing the mean reaction rate as a function of time.

GENERAL CONDITION FOR OPTIMAL ACTIVITY DISTRIBUTION

Consider the general deactivation process described above [eqs (5)–(7) and (9)] in a symmetric domain with boundary conditions (8) and initial condition (10). The goal is to find the initial distribution $\hat{a}(\varphi, 0)$ subject to the constraints

$$(n+1) \int_0^1 \varphi^n \hat{a}(\varphi, 0) \, d\varphi = 1 \text{ and } \hat{a}(\varphi, 0) \geqslant 0 \tag{14}$$

and the time $\hat{\tau} > 0$, such that for $\tau^* = \hat{\tau}$ and $a(\varphi, 0) = \hat{a}(\varphi, 0)$ the objective function (2) is maximized. In the Appendix the following necessary condition for optimality is developed:

If $\hat{a}(\varphi, 0)$ is optimal, then, for any given initial distribution $a(\varphi, 0)$, one has

$$\int_0^1 \varphi^n \Psi(\varphi, 0) \hat{a}(\varphi, 0) \, d\varphi \geqslant \int_0^1 \varphi^n \Psi(\varphi, 0) a(\varphi, 0) \, d\varphi \tag{15}$$

where $\Psi(\varphi, \tau)$ is obtained as a solution of the system of adjoint equations given in the Appendix with coefficients depending on $\hat{a}(\varphi, 0)$, whose detailed form is in fact irrelevant at this stage.

In order to satisfy condition (15), $\hat{a}(\varphi, 0)$ can be "substantially" non-zero solely at points φ_o at which $\Psi(\varphi, 0)$ attains its maximum over [0, 1]. By "substantially non-zero" we mean

$$\int_{\varphi_o - \varepsilon}^{\varphi_o + \varepsilon} \hat{a}(\varphi, 0) \varphi^n \, d\varphi \neq 0 \tag{16}$$

for arbitrarily small $\varepsilon > 0$.

Indeed, suppose this is not true, i.e. that eq. (16) holds while

$$\Psi(\varphi_o', 0) > \Psi(\varphi_o, 0) \tag{17}$$

for some $\varphi_o' \neq \varphi_o$. We show that in such a case it is possible to construct another distribution $a(\varphi, 0)$ satisfying the constraints (11) which violates the optimality criterion (15).

From eq. (17) and the continuity of Ψ it follows that, for sufficiently small $\varepsilon > 0$, all the values of $\Psi(\varphi, 0)$ in the interval $\varphi_o - \varepsilon \leqslant \varphi \leqslant \varphi_o + \varepsilon$ are smaller than any of its value in the interval $\varphi'_o - \varepsilon \leqslant \varphi' \leqslant \varphi'_o + \varepsilon$. Define

$$a(\varphi, 0) = \begin{cases} \hat{a}(\varphi, 0) + (\varphi - \varphi'_o + \varphi_o)^n \hat{a}(\varphi - \varphi'_o + \varphi_o, 0)/\varphi^n & \text{for } \varphi'_o - \varepsilon \leqslant \varphi \leqslant \varphi'_o + \varepsilon \\ 0 & \text{for } \varphi_o - \varepsilon \leqslant \varphi \leqslant \varphi_o + \varepsilon \\ \hat{a}(\varphi, 0) & \text{otherwise.} \end{cases} \quad (18)$$

Then, we have $a(\varphi, 0) \geqslant 0$ and

$$\begin{aligned} \int_0^1 \varphi^n a(\varphi, 0)\,\mathrm{d}\varphi &= \int_{\varphi'_o - \varepsilon}^{\varphi'_o + \varepsilon} \varphi^n \Big[\hat{a}(\varphi, 0) + \frac{(\varphi - \varphi'_o + \varphi_o)^n}{\varphi^n} \hat{a}(\varphi - \varphi'_o + \varphi_o, 0)\Big]\,\mathrm{d}\varphi \\ &+ \int_{\langle 0,1\rangle \setminus (\langle \varphi_o - \varepsilon, \varphi_o + \varepsilon\rangle \cup \langle \varphi'_o - \varepsilon, \varphi'_o + \varepsilon\rangle)} \varphi^n \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &= \int_{\langle 0,1\rangle \setminus \langle \varphi_o - \varepsilon, \varphi_o + \varepsilon\rangle} \varphi^n \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &+ \int_{\varphi'_o - \varepsilon}^{\varphi'_o + \varepsilon} (\varphi - \varphi'_o + \varphi_o)^n \hat{a}(\varphi - \varphi'_o + \varphi_o)\,\mathrm{d}\varphi \\ &= \int_{\langle 0,1\rangle \setminus \langle \varphi_o - \varepsilon, \varphi_o + \varepsilon\rangle} \varphi^n \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &+ \int_{\varphi_o - \varepsilon}^{\varphi_o + \varepsilon} \varphi^n \hat{a}(\varphi, 0)\,\mathrm{d}\varphi = \int_0^1 \varphi^n \hat{a}(\varphi, 0)\,\mathrm{d}\varphi = 1. \end{aligned} \quad (19)$$

Hence $a(\varphi, 0)$ satisfies constraints (11).

Since from eq. (17) it follows that $\Psi(\varphi - \varphi_o + \varphi'_o, 0) > \Psi(\varphi, 0)$ for all $\varphi_o - \varepsilon \leqslant \varphi \leqslant \varphi_o + \varepsilon$, through similar manipulations we obtain

$$\begin{aligned} &\int_0^1 \varphi^n \Psi(\varphi, 0)\, a(\varphi, 0)\,\mathrm{d}\varphi \\ &= \int_{\langle 0,1\rangle \setminus \langle \varphi_o - \varepsilon, \varphi_o + \varepsilon\rangle} \varphi^n \Psi(\varphi, 0)\, \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &+ \int_{\varphi_o - \varepsilon}^{\varphi_o + \varepsilon} \Psi(\varphi - \varphi_o + \varphi'_o, 0)\, \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &> \int_{\langle 0,1\rangle \setminus \langle \varphi_o - \varepsilon, \varphi_o + \varepsilon\rangle} \varphi^n \Psi(\varphi, 0)\, \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &+ \int_{\varphi_o - \varepsilon}^{\varphi_o + \varepsilon} \varphi^n \Psi(\varphi, 0)\, \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \\ &= \int_0^1 \varphi^n \Psi(\varphi, 0)\, \hat{a}(\varphi, 0)\,\mathrm{d}\varphi \end{aligned} \quad (20)$$

which contradicts the necessary condition for optimality (15).

As a conclusion, it can be observed that the optimality criterion (15) practically excludes any initial distribution $a(\varphi, 0)$ which is not of the Dirac-delta type. Indeed, it is highly unlikely that $\Psi(\varphi, 0)$ would attain its maximum simultaneously at more than one point at which $a(\varphi, 0)$ is substantially non-zero and therefore it is not meaningful to consider this possibility any further. Consequently, having found the optimal distribution in the class of all one-peak Dirac-delta ones we do not expect any other distribution to improve the objective functional any further.

In fact, for any given distribution, we have shown through eq. (15) that it is possible to construct a suitable Dirac delta distribution which improves the objective functional (2). Of course, the mathematical proof presented above is not completely rigorous and we are not very optimistic about the change for it to be found since $\Psi(\varphi, 0)$ is obtained as a result of solving a system of partial differential equations, the coefficients of which depend on $\hat{a}(\varphi, 0)$ itself. However, since $\Psi(\varphi, 0)$ is the result of an integration process, it certainly has some continuity and well-behaving properties. For this reason, condition (15) can be regarded as a fully satisfactory mathematical justification to consider one-peak Dirac-delta function initial distributions as the only candidates for optimal ones.

In addition, as we will see below, criterion (15) may exclude some delta distributions as well and indicate in which direction to move the activity location point to find the optimal one.

APPLICATION TO THE CASE OF INDEPENDENT POISONING WITH FIRST-ORDER ISOTHERMAL REACTIONS

Once it has been established that the optimal activity distribution is of the Dirac-delta type, the optimization problem reduces to the selection, among all possible Dirac-delta distributions, of the optimal one depending upon the particular reacting system and operating conditions under consideration. For illustrative purposes, let us consider the case of an isothermal first-order reaction with dimensionless rate equation

$$R = aY \quad (21)$$

which occurs together with independent chemisorption of catalyst poison, leading to the following rate expression for the deactivation process:

$$R_p = aY_p. \quad (22)$$

First, we consider a Dirac-delta activity distribution located at the point φ_1. The initial condition, in coincidence with eq. (11), is

$$a(\varphi, 0) = \frac{\delta(\varphi - \varphi_1)}{(n+1)\varphi_1^n} \quad \text{at } \tau = 0. \quad (23)$$

For $\tau > 0$ it is convenient to express the activity as a product of the initial activity distribution and

a time-dependent variable $\mu(\tau)$, i.e.

$$a(\varphi, \tau) = a(\varphi, 0)\,\mu(\tau) \tag{24}$$

where $\mu(0) = 1$.

The solution of the model eqs (5), (6) with the rate expressions (21) and (22) and with boundary conditions (8) can be written in the following closed form:

$$\varphi \in \langle 0, \varphi_1): \; Y = Y_1, \; Y_p = Y_{p1} \tag{25}$$

$$\varphi = \varphi_1: \; Y = Y_1 = (n+1)/(n+1+x\mu) \tag{26}$$

$$Y_p = Y_{p1} = (n+1)/(n+1+\alpha x\mu) \tag{27}$$

$$\varphi \in (\varphi_1, 1\rangle: \; Y = 1 - (1 - Y_1)\zeta_n(\varphi)/\zeta_n(\varphi_1) \tag{28}$$

$$Y_p = 1 - (1 - Y_{p1})\zeta_n(\varphi)/\zeta_n(\varphi_1) \tag{29}$$

where
$$\zeta_n(\varphi) = \begin{cases} 1 - \varphi & \text{for } n = 0 \\ \ln(1/\varphi) & \text{for } n = 1 \\ (1-\varphi)/\varphi & \text{for } n = 2 \end{cases} \tag{30}$$

and $x = \Phi^2\zeta_n(\varphi_1)$, and $\alpha = \Phi_p^2/\Phi^2$. Substituting eqs (23), (24), (27) and the deactivation rate expression (22) reduces eq. (9) to

$$\frac{\partial a}{\partial \tau} = -\frac{\delta(\varphi - \varphi_1)}{(n+1)\varphi_1^n}\mu\frac{n+1}{n+1+\alpha x\mu} = \frac{\delta(\varphi - \varphi_1)}{(n+1)\varphi_1^n}\frac{\mathrm{d}\mu}{\mathrm{d}\tau} \tag{31}$$

which leads to

$$\frac{\mathrm{d}\mu}{\mathrm{d}\tau} = -\mu\frac{n+1}{n+1+\alpha x\mu}. \tag{32}$$

Integrating from $\tau = 0$ to τ and $\mu = 1$ to μ we obtain

$$\tau = \frac{\alpha x}{n+1}(1-\mu) - \ln\mu. \tag{33}$$

When using the Dirac-delta activity distribution (23), the objective functional (2) becomes a function of two parameters: the active point location φ_1 and the operating time τ^*. Equation (33) relating τ and μ allows us to express the objective function (2) in a closed form as a function of the active point location φ_1 and the relative activity μ^* at time τ^*:

$$\mathscr{J}(\varphi_1, \mu^*) = \frac{\gamma\alpha(n+1)\left(\frac{n+1}{x}\frac{\alpha-1}{\alpha}\ln\frac{n+1+x\mu^*}{n+1+x} + 1 - \mu^*\right) - (n+1)}{\alpha x(1-\mu^*) - (n+1)\ln\mu^*} \tag{34}$$

For $0 < \alpha \leqslant 1$ it readily appears that for $x \geqslant 0$ the denominator of eq. (34) is increasing while the numerator is decreasing. It follows that, for fixed μ^*, in those intervals of x on which $\mathscr{J}$ is positive, it decreases with x. Consequently, for those values of μ^* for which $\mathscr{J}$ is positive for $x = 0$ it attains its maximum with respect to $x \geqslant 0$ at this point. Since $x = 0$ corresponds to $\varphi_1 = 1$ this means that all the active catalyst should be located at the external pellet surface. This is because in this case $\Phi > \Phi_p$, i.e. the intraparticle transport resistance is larger for the main reactant than for the poison.

Note that if for some μ^* the value of $\mathscr{J}$ is negative for $x = 0$ then it remains negative for all $x \geqslant 0$. Such values of μ^* are uninteresting since for them the catalytic process cannot be profitable no matter where the active catalyst is placed.

To find the optimal value of μ^* we have to maximize the function $z(\mu^*) = \mathscr{J}|_{x=0}$ in the subinterval of $\langle 0, 1\rangle$, in which z is non-negative.

Using de l'Hospital's rule we obtain a formula for z which is independent of α:

$$z(\mu^*) = [(\mu^* - 1)\gamma + 1]/\ln\mu^*. \tag{35}$$

In order to have $z(\mu^*) \geqslant 0$ for some $0 \leqslant \mu^* \leqslant 1$ one needs $\gamma \geqslant 1/(1-\mu^*)$: for such γ, $z(\mu^*) \geqslant 0$ for $0 \leqslant \mu^* \leqslant 1 - 1/\gamma$. Note that from the expression for the right boundary of this interval γ can be expressed as

$$\gamma = 1/(1-\mu^*). \tag{36}$$

To maximize $z(\mu^*)$ we first find its local extrema. Those are the solutions of the equation $z'(\mu^*) = 0$ which is

$$\gamma\ln\mu^* + \gamma(1-\mu^*)/\mu^* - 1/\mu^* = 0. \tag{37}$$

From this equation we can express γ as a function of μ^*:

$$\gamma = 1/(\mu^*\ln\mu^* + 1 - \mu^*). \tag{38}$$

Since $\ln\mu^* < 0$ and $\mu^*\ln\mu^* + 1 - \mu^* > 0$ for all $0 \leqslant \mu^* \leqslant 1$ from eq. (37) it follows that

$$\gamma \geqslant 1/(1-\mu^*). \tag{39}$$

Comparing eqs (36) and (39) we see that for all $\gamma > 1/(1-\mu^*)$ there is a unique root of eq. (37) in the interval $\langle 0, 1 - 1/\gamma\rangle$: this root is the optimal value of μ^*.

We can summarize our analysis as follows. For each $0 < \alpha \leqslant 1$ there is a positive threshold value of γ below which the process cannot be profitable for any choice of τ^* and φ_1. For γ above this threshold value the optimal location of the active catalyst is always at the boundary of the pellet and the optimal operating time moves monotonically from infinity to zero for the parameter γ moving from the threshold value to infinity.

Since for $\alpha > 1$ both the numerator and the denominator of eq. (34) increase with x the optimal location of the active catalyst can be somewhere inside the pellet. Its precise location depends upon the parameter γ, the reaction kinetic parameters and the pellet geometry. In this case the maximum of function (34) with respect to x (i.e. φ_1) and μ^* have to be found numerically, using any of the standard optimization techniques available in the literature. As an example,

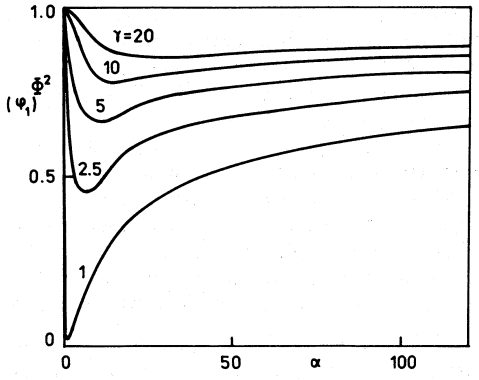


Fig. 1. $(\varphi_1)^{\Phi^2}$ vs α for various values of γ ($n = 1$).

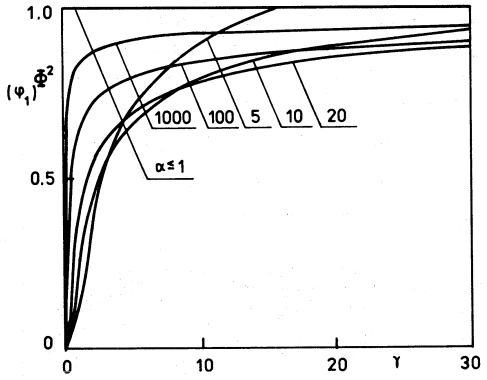


Fig. 2. $(\varphi_1)^{\Phi^2}$ vs γ for various values of α ($n = 1$).

the effect of the parameters γ and α on the optimal active catalyst location for a cylindrical pellet ($n = 1$) is shown in Figs 1 and 2, respectively. The corresponding values of the optimal operating time, optimal residual activity and optimal objective function are shown in Figs 3–5. From the results shown in Figs 1 and 3 it appears that, for fixed γ and increasing

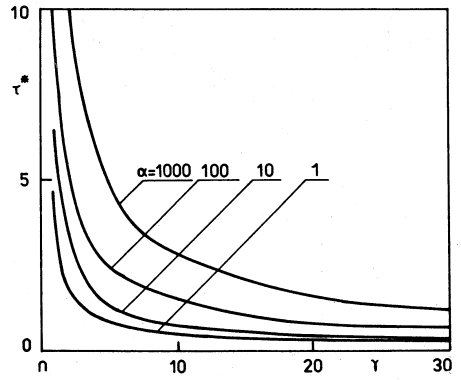


Fig. 3. Optimal operation time, τ^* vs γ, for various values of α ($n = 1$).

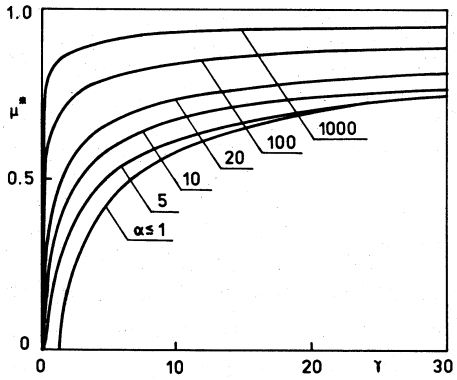


Fig. 4. Optimal residual activity, μ^* vs γ, for various values of α ($n = 1$).

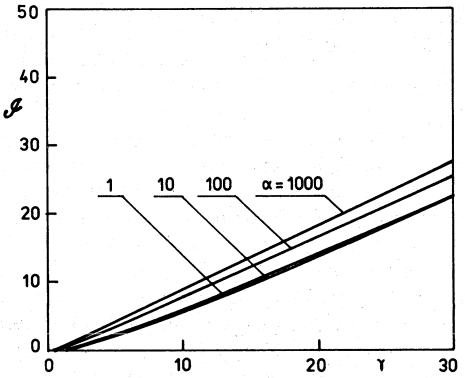


Fig. 5. Objective function, $\mathscr{J}$ vs γ, for various values of α ($n = 1$).

values of the ratio $\alpha = \Phi_p^2/\Phi^2$, the optimal location first moves towards the pellet interior but then comes back towards the pellet external surface, while the optimal operation time increases monotonically. On the other hand, from Figs 2 and 3 it appears that, for fixed α, decreasing values of the catalyst replacement cost (i.e. increasing values of γ) lead to optimal catalyst locations closer to the pellet external surface and to lower optimal operating times, τ^*. However, as expected, from the results shown in Fig. 5 it appears that better performance is achieved when the intraparticle diffusion resistance of the poison is larger than that of the main reactant.

In order to further support the results of the theoretical analysis reported above, and to investigate the possibility of their transfer into practical applications, let us consider the following step distribution function

$$\begin{aligned} &\varphi \in \langle 0, \varphi_1) \text{ and } \varphi \in (\varphi_2, 1\rangle\colon a(\varphi, \tau) = 0 \\ &\varphi \in \langle \varphi_1, \varphi_2\rangle\colon a(\varphi, 0) = 1/(\varphi_2^{n+1} - \varphi_1^{n+1}) \end{aligned} \tag{40}$$

which is such as to satisfy constraint (11). By recalling that the reaction does not take place outside the interval $\langle \varphi_1, \varphi_2\rangle$ the pellet mass balances can be

recasted in the following form:

$$\varphi \in \langle 0, \varphi_1): \; Y = Y_1, \; Y_p = Y_{p1} \tag{41}$$

$$\varphi = \varphi_1: \; Y = Y_1, \; Y_p = Y_{p1}$$

$$\partial Y/\partial\varphi = \partial Y_p/\partial\varphi = 0 \tag{42}$$

$$\varphi \in \langle \varphi_1, \varphi_2 \rangle: \; \nabla^2 Y = \Phi^2 R, \; \nabla^2 Y_p = \Phi_p^2 R_p \tag{43}$$

$$\varphi = \varphi_2: \; Y = Y_2, \; Y_p = Y_{p2}$$

$$\partial Y/\partial\varphi = -(1 - Y_2)(\mathrm{d}\zeta_n/\mathrm{d}\varphi)_{\varphi=\varphi_2}/\zeta_n(\varphi_2)$$

$$\partial Y_p/\partial\varphi = -(1 - Y_{p2})(\mathrm{d}\zeta_n/\mathrm{d}\varphi)_{\varphi=\varphi_2}/\zeta_n(\varphi_2) \tag{44}$$

$$\varphi \in (\varphi_2, 1\rangle: \; Y = 1 - (1 - Y_2)\zeta_n(\varphi)/\zeta_n(\varphi_2)$$

$$Y_p = 1 - (1 - Y_{p2})\zeta_n(\varphi)/\zeta_n(\varphi_2). \tag{45}$$

The system of eqs (41)–(45) has been solved numerically, by discretizing the space coordinate φ by a standard finite-difference scheme. The activity value at each position inside the pellet has been computed as a function of time by integrating eq. (9) by a marching technique. The values of the objective function (2) and of the effectiveness factor (13) have been computed through suitable quadrature formulae.

In the case of step activity distribution, the objective function (2) is a function of three parameters, φ_1, φ_2 and τ^*. In all the performed optimization runs the optimal initial activity distribution has been given by the narrowest possible step distribution of the adopted discretization procedure (i.e. $\varphi_2 - \varphi_1$ = step size in the finite-difference scheme) centered at the optimal location predicted by the ideal Dirac-delta distribution. A typical example is shown in Fig. 6, where the curves represent step distributions (i.e. values of φ_1 and φ_2 with optimized operating time, τ^*) which exhibit the same value of the objective function $\mathcal{J}$ (2). It clearly appears that the best performance corresponds to a Dirac-delta activity distribution, i.e. $\varphi_1 = \varphi_2 = 0.67$. It is also worth noting that the performance of such a Dirac-delta distribution is actually quite closely approached by step distribution of relatively small width and centered at the same point. This result is of great importance with respect to the actual preparation of such optimally distributed catalyst pellets, as it has been previously discussed in the context of non-deactivating reacting systems (Morbidelli *et al.*, 1982).

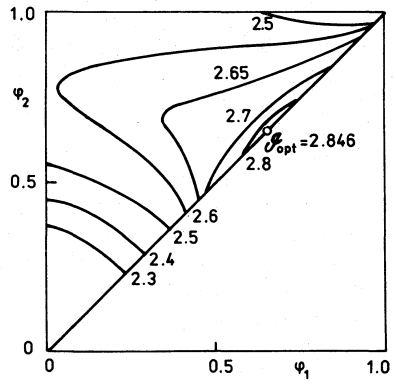


Fig. 6. Level lines of the objective function $\mathcal{J}$ for various step activity distributions (parameter values: $\alpha = 10$, $\gamma = 5$, $\Phi^2 = 1$, $n = 1$).

Finally, in order to further illustrate the general condition for optimality (15) the adjoint variable profiles $\Psi(\varphi, 0)$ for the reacting systems under examination, are shown in Fig. 7(a) and (b) for step and Dirac-delta activity distributions, respectively. These have been obtained by solving numerically the system of the adjoint equations which in the case under examination reduces to

$$\nabla^2 p + a(1 - p\Phi^2) = 0 \tag{46}$$

$$\nabla^2 q - q\Phi_p^2 a - \Psi a = 0 \tag{47}$$

$$\partial\Psi/\partial\tau + Y(1 - p\Phi^2) - q\Phi_p^2 Y_p - \Psi Y_p = 0 \tag{48}$$

with boundary and terminal conditions

$$\tau = \hat{\tau}: \; \Psi(\varphi, \tau) = 0 \tag{49}$$

$$\tau \in \langle 0, \hat{\tau}\rangle; \; \varphi = 0: \; \partial p/\partial\varphi = \partial q/\partial\varphi = \partial\Psi/\partial\varphi = 0 \tag{50}$$

$$\varphi = 1: \; p = q = 0. \tag{51}$$

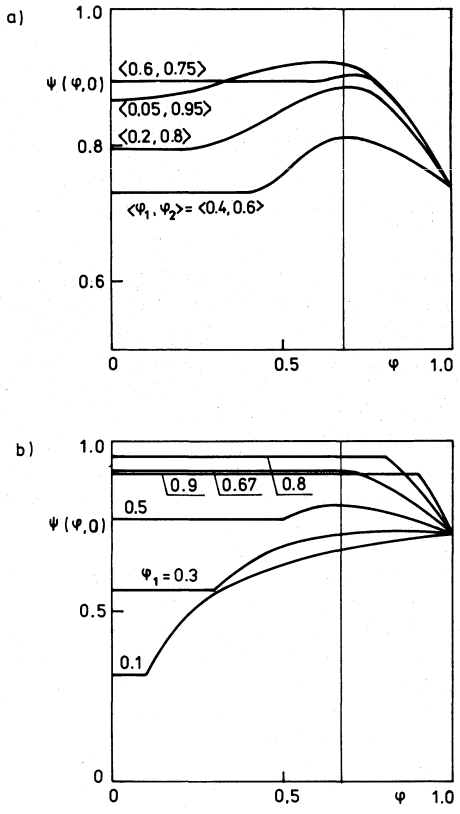


Fig. 7. Adjoint variable $\Psi(\varphi, 0)$ profiles. The vertical solid line indicates location of the optimal Dirac-delta distribution (parameter values: $\alpha = 10$, $\gamma = 5$, $\Phi^2 = 1$, $n = 1$). (a) Step function activity distributions. (b) Dirac-delta activity distributions.

Since the system parameters adopted in Figs 6 and 7 are the same, it can be noted that also in the latter case is the optimal Dirac-delta distribution located at $\varphi_1 = 0.67$ [as indicated by the solid vertical line in Fig. 7(a) and (b)]. It is rather surprising that for all step-size distributions considered, even the widest one $\langle 0.05, 0.95\rangle$, the maximum of the $\Psi(\varphi, 0)$ curves is very close to the location of the optimal Dirac-delta distribution. This provides a useful initial information for the optimum search. In Fig. 7(b) the adjoint profiles $\Psi(\varphi, 0)$ are shown relative to Dirac-delta distributions centered at various locations φ_1. It appears that using criterion (15) it is possible to exclude the locations $\varphi_1 = 0.1$, 0.3 and 0.5, since the corresponding adjoint functions exhibit their maximum values at other locations. In addition, the function $\Psi(\varphi, 0)$ indicates in all cases that the optimal location should be to the right (i.e. larger values of φ_1) since the value of the integral in the right-hand side of condition (15) increases when moving the Dirac-delta location in this direction. On the other hand, criterion (15) is not fine enough to exclude the location points to the right of the optimal one (i.e. 0.8 and 0.9). The only way to exclude such points is in fact by comparing the corresponding values of the objective function, as has been done in the numerical optimization procedure described above.

CONCLUDING REMARKS

A method for determining the optimal activity distribution in catalyst pellets for reacting systems undergoing deactivation has been developed. As the objective function, profit per time taking into account the price of the product and the cost of catalyst replacement or regeneration has been considered. A general condition for optimality has been developed which allows one to conclude that Dirac-delta activity distributions are the only candidates for optimal ones. It is remarkable that such a conclusion is of quite general validity, since it applies to any kind of kinetic expression for the main reaction as well as for the poisoning process and any pellet geometry. Even though not reported here in detail for brevity reasons, it is worth mentioning that the same conclusion can be reached when accounting for external mass and heat transfer resistances as well as for other types of objective functions. In such cases the derivation follows closely the arguments reported in the recent paper by Wu *et al.* (1990) referring to the case of non-isothermal reacting systems of a fully general nature in the absence of deactivation. Also in this case it has been found that the optimal activity distribution is of the Dirac delta type.

As an illustrative example the case of an isothermal first-order reaction with independent poisoning has been investigated. The effect of the kinetic parameters and operating conditions on the optimal location of the Dirac-delta distribution, as well as on the optimal operating time, has been discussed in detail. In particular, it has been found that for values of the ratio between the poison and the main-reactant Thiele moduli smaller than one the optimal location of the active catalyst is at the external pellet surface. For values of this parameter larger than one the optimal location moves towards the pellet interior to an extent which depends upon the specific operating conditions under examination. Finally, in order to establish the possibility of transferring the results of this work to practical applications, the performance of step activity distributions has been investigated. Such distributions have in fact been investigated experimentally in the context of optimal catalyst design for non-deactivating systems [cf. Wu *et al.* (1990)].

NOTATION

a	activity
$\boldsymbol{a}$	characteristic dimension of catalyst pellet
a_p	equilibrium poison adsorbed amount
C	concentration
D	diffusion coefficient
$(-\Delta H)$	heat of reaction
$\mathscr{J}$	objective function
k	reaction rate constant
n	integer characteristic of pellet geometry ($n = 0$, slab; $n = 1$, cylinder; $n = 2$, sphere)
p	adjoint variable
q	adjoint variable
r	reaction rate
R	dimensionless main reaction rate
R_P	dimensionless poisoning rate
s	adjoint variable
t	time
T	temperature
t^o	characteristic deactivation time
x	$= \Phi^2 \zeta_n(\varphi_1)$, dimensionless parameter
Y	$= C_A/C_{Ao}$, dimensionless reactant concentration
Y_p	$= C_p/C_{po}$, dimensionless poison concentration

Greek letters

α	$= \Phi_p^2/\Phi^2$, ratio of Thiele moduli
α_1	price of product
α_2	cost of catalyst
β	$= (-\Delta H) D_A C_{Ao}/(\lambda T_o)$, dimensionless reaction heat
γ	$= \alpha_1/\alpha_2$, dimensionless parameter
η	effectiveness factor
λ	thermal conductivity
μ	relative activity defined by eq. (24)
σ	concentration of available catalytically active sites
τ	$= t/t^o$, dimensionless time
υ	$= T/T_o$, dimensionless temperature
φ	dimensionless space coordinate
Φ	$= a[r_o/(D_A C_{Ao})]^{1/2}$, reaction Thiele modulus
Φ_p	$= a[a_p/(D_p C_{po} t^o)]^{1/2}$, poison Thiele modulus
Ψ	adjoint variable

Subscripts

*	terminal conditions
o	surface conditions
p	poison
1, 2	activity location

REFERENCES

Becker, E. and Wei, J., 1977a, Nonuniform distribution of catalysts on supports. I. Bimolecular Langmuir reactions. *J. Catal.* **46**, 365–371.

Becker, E. and Wei, J., 1977b, Nonuniform distribution of catalysts on supports. II. First order reactions with poisoning. *J. Catal.* **46**, 372–381.

Chemburkar, R. M., Morbidelli, M. and Varma, A., 1987, Optimal catalyst activity profiles in pellets—VII. The case of arbitrary reaction kinetics with finite external heat and mass transport resistances. *Chem. Engng Sci.* **42**, 2621–2632.

Corbett, W. E. and Luss, D., 1974, The influence of nonuniform catalytic activity on the performance of a single spherical pellet. *Chem. Engng Sci.* **29**, 1473–1483.

De Lancey, G. B., 1973, An optimal catalyst activation policy for poisoning problems. *Chem. Engng Sci.* **28**, 105–118.

Dougherty, R. C. and Verykios, X. E., 1987, Nonuniformly activated catalysts. *Catal. Rev. Sci. Engng* **29**, 101–150.

Morbidelli, M., Servida, A. and Varma, A., 1982, Optimal catalyst activity profiles in pellets—I. The case of negligible external mass transfer resistance. *Ind. Engng Chem. Fundam.* **21**, 278–284.

Morbidelli, M., Servida, A., Carrà, S. and Varma, A., 1985, Optimal activity profiles in pellets—III. *Ind. Engng Chem. Fundam.* **24**, 116–118.

Wu, H., Brunovská, A., Morbidelli, M. and Varma, A., 1990, Optimal catalyst activity profiles in pellets—IX. General nonisothermal reacting systems with arbitrary kinetics. *Chem. Engng Sci.* (in press).

APPENDIX: DEVELOPMENT OF THE OPTIMALITY CONDITION (15)

Let $\hat{\tau}$ be the optimal time and let $\hat{a}(\varphi, 0)$ be the optimal initial activity distribution. Then, it follows that

$$I(\hat{a}) = \int_0^{\hat{\tau}} \eta \, d\tau = (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n R \, d\varphi \, d\tau \tag{A1}$$

is maximum over all $\hat{a}$s subject to the constraints (14). In particular, if we take any $a(\varphi, 0)$ satisfying eq. (14), then

$$a_\varepsilon(\varphi, 0) = \hat{a}(\varphi, 0) + \varepsilon[a(\varphi, 0) - \hat{a}(\varphi, 0)] \tag{A2}$$

will satisfy eq. (14) for all $0 \leqslant \varepsilon \leqslant 1$. Thus, in order to satisfy eq. (A1) we must have

$$\frac{d}{d\varepsilon} I(a_\varepsilon)|_{\varepsilon=0} \leqslant 0. \tag{A3}$$

Let us denote $\delta X = dX/d\varepsilon$. By differentiating eqs (5)–(7), (9), the boundary condition (8) and eq. (A2) with respect to ε at $\varepsilon = 0$ we obtain

$$\nabla^2 \delta Y = \Phi^2 \left(\frac{\partial R}{\partial Y} \partial Y + \frac{\partial R}{\partial Y_p} \delta Y_p + \frac{\partial R}{\partial \upsilon} \delta \upsilon + \frac{\partial R}{\partial a} \partial a \right) \tag{A4}$$

$$\nabla^2 \delta Y_p = \Phi_p^2 \left(\frac{\partial R_p}{\partial Y} \delta Y + \frac{\partial R_p}{\partial Y_p} \delta Y_p + \frac{\partial R_p}{\partial \upsilon} \delta \upsilon + \frac{\partial R_p}{\partial a} \delta a \right) \tag{A5}$$

$$\nabla^2 \delta \upsilon = -\beta \Phi^2 \left(\frac{\partial R}{\partial Y} \partial Y + \frac{\partial R}{\partial Y_p} \delta Y_p + \frac{\partial R}{\partial \upsilon} \delta \upsilon + \frac{\partial R}{\partial a} \delta a \right) \tag{A6}$$

$$-\frac{\partial \delta a}{\partial \tau} = \frac{\partial R_p}{\partial Y} \delta Y + \frac{\partial R_p}{\partial Y_p} \delta Y_p + \frac{\partial R_p}{\partial \upsilon} \delta \upsilon + \frac{\partial R_p}{\partial a} \delta a \tag{A7}$$

$$\varphi = 1: \delta Y(1, \tau) = \delta Y_p(1, \tau) = \delta \upsilon(1, \tau) = 0 \tag{A8}$$

$$\tau = 0: \delta a(\varphi, 0) = a(\varphi, 0) - \hat{a}(\varphi, 0). \tag{A9}$$

Thus using eqs (A4)–(A6), eq. (A1) reduces to

$$\begin{aligned}
\delta I = {} & (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \left(\frac{\partial R}{\partial Y} \delta Y + \frac{\partial R}{\partial Y_p} \delta Y_p \right. \\
& \left. + \frac{\partial R}{\partial \upsilon} \delta \upsilon + \frac{\partial R}{\partial a} \delta a \right) d\varphi \, d\tau \\
& + (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n p \left[\nabla^2 \delta Y - \Phi^2 \left(\frac{\partial R}{\partial Y} \delta Y \right. \right. \\
& \left. \left. + \frac{\partial R}{\partial Y_p} \delta Y_p \; \frac{\partial R}{\partial \upsilon} \delta \upsilon + \frac{\partial R}{\partial a} \delta a \right) \right] d\varphi \, d\tau \\
& + (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n q \left[\nabla^2 \delta Y_p - \Phi_p^2 \left(\frac{\partial R_p}{\partial Y} \delta Y \right. \right. \\
& \left. \left. + \frac{\partial R_p}{\partial Y_p} \delta Y_p + \frac{\partial R_p}{\partial \upsilon} \delta \upsilon + \frac{\partial R_p}{\partial a} \delta a \right) \right] d\varphi \, d\tau \\
& + (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n s \left[\nabla^2 \delta \upsilon + \beta \Phi^2 \left(\frac{\partial R}{\partial Y} \delta Y \right. \right. \\
& \left. \left. + \frac{\partial R}{\partial Y_p} \delta Y_p + \frac{\partial R}{\partial \upsilon} \delta \upsilon + \frac{\partial R}{\partial a} \delta a \right) \right] d\varphi \, d\tau \\
& + (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \frac{\partial}{\partial \tau} (\Psi \delta a) \, d\varphi \, d\tau \\
& + (n+1) \int_0^1 \varphi^n \Psi(\varphi, 0) \, \delta a(\varphi, 0) \, d\varphi
\end{aligned} \tag{A10}$$

where $p(\varphi, \tau)$, $q(\varphi, \tau)$, $s(\varphi, \tau)$ and $\Psi(\varphi, \tau)$ are adjoint variables (corresponding to Lagrangian multipliers) and the following relationships have been used:

$$\Psi(\varphi, \hat{\tau}) = 0 \tag{A11}$$

$$\begin{aligned}
& (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \frac{\partial}{\partial \tau} (\Psi \delta a) \, d\varphi \, d\tau \\
& = (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \left[\frac{\partial \Psi}{\partial \tau} \delta a + \Psi \frac{\partial \delta a}{\partial \tau} \right] d\varphi \, d\tau \\
& = (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \left[\frac{\partial \Psi}{\partial \tau} \delta a - \Psi \left(\frac{\partial R_p}{\partial Y} \delta Y + \frac{\partial R_p}{\partial Y_p} \delta Y_p \right. \right. \\
& \left. \left. + \frac{\partial R_p}{\partial \upsilon} \partial \upsilon + \frac{\partial R_p}{\partial a} \delta a \right) \right] d\varphi \, d\tau
\end{aligned} \tag{A12}$$

By selecting p, q and s so as to satisfy the boundary conditions

$$\varphi = 1: p(1, \tau) = q(1, \tau) = s(1, \tau) = 0 \tag{A13}$$

and integrating by parts twice one obtains

$$\int_0^1 \varphi^n p \nabla^2 \delta Y \, d\varphi = \int_0^1 \varphi^n \delta Y \; \nabla^2 p \, d\varphi \tag{A14}$$

$$\int_0^1 \varphi^n q \nabla^2 \delta Y_p \, d\varphi = \int_0^1 \varphi^n \delta Y_p \; \nabla^2 q \, d\varphi \tag{A15}$$

$$\int_0^1 \varphi^n s \nabla^2 \delta \upsilon \, d\varphi = \int_0^1 \varphi^n \delta \upsilon \; \nabla^2 s \, d\varphi. \tag{A16}$$

Substituting eqs (A12) and (A14)–(A16) into eq. (A10) and grouping terms multiplied by δY, δY_p, $\delta \upsilon$ and δa into separate integrals one obtains

$$\begin{aligned}
\delta I = {} & (n+1) \int_0^{\hat{\tau}} \int_0^1 \varphi^n \delta Y \left(\frac{\partial R}{\partial Y} + \nabla^2 p - p \Phi^2 \frac{\partial R}{\partial Y} \right. \\
& \left. - q \Phi_p^2 \frac{\partial R_p}{\partial Y} + s \beta \Phi^2 \frac{\partial R}{\partial Y} - \Psi \frac{\partial R_p}{\partial Y} \right) d\varphi \, d\tau
\end{aligned}$$

$$+ (n+1)\int_0^{\tau}\int_0^1 \varphi^n \delta Y_p \left(\frac{\partial R}{\partial Y_p} - p\Phi^2 \frac{\partial R}{\partial Y_p} + \nabla^2 q - q\Phi_p^2 \frac{\partial R_p}{\partial Y_p} + s\beta\Phi^2 \frac{\partial R}{\partial Y_p} - \Psi \frac{\partial R_p}{\partial Y_p}\right) \mathrm{d}\varphi\, \mathrm{d}\tau$$

$$+ (n+1)\int_0^{\tau}\int_0^1 \varphi^n \delta\upsilon \left(\frac{\partial R}{\partial \upsilon} - p\Phi^2 \frac{\partial R}{\partial \upsilon} - q\Phi_p^2 \frac{\partial R_p}{\partial \upsilon} + \nabla^2 s + s\beta\Phi^2 \frac{\partial R}{\partial \upsilon} - \Psi \frac{\partial R_p}{\partial \upsilon}\right) \mathrm{d}\varphi\, \mathrm{d}\tau$$

$$+ (n+1)\int_0^{\tau}\int_0^1 \varphi^n \delta a \left(\frac{\partial R}{\partial a} - p\Phi^2 \frac{\partial R}{\partial a} - q\Phi_p^2 \frac{\partial R_p}{\partial a} + s\beta\Phi^2 \frac{\partial R}{\partial a} - \Psi \frac{\partial R_p}{\partial a} + \frac{\partial \Psi}{\partial \tau}\right) \mathrm{d}\varphi\, \mathrm{d}\tau$$

$$+ (n+1)\int_0^1 \varphi^n \Psi(\varphi, 0)\, \delta a(\varphi, 0)\, \mathrm{d}\varphi. \qquad \text{(A17)}$$

By selecting p, q, s and Ψ so as to satisfy the following system of adjoint equations:

$$\nabla^2 p + \frac{\partial R}{\partial Y}(1 - p\Phi^2 + s\beta\Phi^2) - \frac{\partial R_p}{\partial Y}(q\Phi_p^2 + \Psi) = 0 \qquad \text{(A18)}$$

$$\nabla^2 q + \frac{\partial R}{\partial Y_p}(1 - p\Phi^2 + s\beta\Phi^2) - \frac{\partial R_p}{\partial Y_p}(q\Phi_p^2 + \Psi) = 0 \qquad \text{(A19)}$$

$$\nabla^2 s + \frac{\partial R}{\partial \upsilon}(1 - p\Phi^2 + s\beta\Phi^2) - \frac{\partial R_p}{\partial \upsilon}(q\Phi_p^2 + \Psi) = 0 \qquad \text{(A20)}$$

$$\frac{\partial \Psi}{\partial \tau} + \frac{\partial R}{\partial a}(1 - p\Phi^2 + s\beta\Phi^2) - \frac{\partial R_p}{\partial a}(q\Phi_p^2 + \Psi) = 0 \qquad \text{(A21)}$$

eq. (A17) reduces to

$$\delta I = (n+1)\int_0^1 \varphi^n \Psi(\varphi, 0)\, \delta a(\varphi, 0)\, \mathrm{d}\varphi = \frac{\mathrm{d}I}{\mathrm{d}\varepsilon}. \qquad \text{(A22)}$$

Finally, using eq. (A9), eq. (A22) leads to

$$\delta I = (n+1)\int_0^1 \varphi^n \Psi(\varphi, 0)\, [a(\varphi, 0) - \hat{a}\,(\varphi, 0)]\, \mathrm{d}\varphi \qquad \text{(A23)}$$

which substituted into eq. (A3) leads to the general condition for optimality

$$\int_0^1 \varphi^n \Psi(\varphi, 0)\, \hat{a}\,(\varphi, 0)\, \mathrm{d}\varphi \geqslant \int_0^1 \varphi^n \psi(\varphi, 0)\, a(\varphi, 0)\, \mathrm{d}\varphi. \qquad \text{(A24)}$$

P. Brunovský, D. Ševčovič

Explanation of spurt for a non-Newtonian fluid by a diffusion term

Quart. Appl. Math. 52(3) (1994), 401–426.

QUARTERLY OF APPLIED MATHEMATICS

VOLUME LII SEPTEMBER · 1994 NUMBER 3

SEPTEMBER 1994, PAGES 401–426

EXPLANATION OF SPURT FOR A NON-NEWTONIAN FLUID BY A DIFFUSION TERM

BY

P. BRUNOVSKÝ AND D. ŠEVČOVIČ

Institute of Applied Mathematics, Comenius University, Czechoslovakia

1. Introduction. A surprising feature of the flow of polymers is associated with a sudden increase in the volumetric flow rate when the pressure gradient is gradually increased beyond a critical value. This striking phenomenon, called "spurt", was apparently first observed by Vinogradov et al. [15] in rheological experiments involving the flow through thin capillaries of highly elastic and very viscous non-Newtonian fluids like some synthesized polybutadienes and polyisoprenes. The interested reader is referred to [15, Table 1] for more detailed information about microstructure characteristics of samples. The spurt phenomenon is a kind of a flow instability in pressure-driven shear flows of viscoelastic fluids.

Much effort is being spent to explain spurt and related phenomena mathematically. Several authors have considered mathematical models based on differential constitutive equations due to Johnson, Sagelman, and Oldroyd exhibiting local extrema of the steady shear stress as a function of steady strain rate (see [6–8, 10–13]). These papers show that the spurt phenomenon is dynamic and, hence, cannot be explained in a satisfactory manner by only studying the steady-state equations. Dynamical theory can explain phenomena observed in experiments and in numerical simulations, and it can also predict phenomena like latency, shape memory, and hysteresis which should be observable in future experiments.

In this paper we modify the models of [6] and [13] by adding a diffusion term to the constitutive equation. The resulting system of equations (in dimensionless units) governing planar shear flow has the form

$$\begin{aligned} \alpha v_t &= v_{xx} + \sigma_x + f, \\ \sigma_t &= -\sigma + g(v_x) + \nu^2 \sigma_{xx} \end{aligned} \tag{1.1}$$

where $v(t, x)$ is the velocity of the planar flow, $\sigma(t, x)$ is the polymer contribution to the shear stress, $g: \Re \to \Re$ is a given smooth function, and $f > 0$ is the pressure gradient driving the flow.

Unlike the models investigated in [13] and [6] and the other models in [10–12], system (1.1) contains the spatial diffusion term $\nu^2 \sigma_{xx}$. Spatial diffusion is usually

Received April 15, 1991.

1991 *Mathematics Subject Classification.* Primary 76A10; Secondary 35B40, 35B25, 34B15.

neglected in non-Newtonian models because of the spatial homogeneity of the structure. In the model of [4] (also see [3]), Brownian motion prevents polymer molecules (treated as dumb-bells) from being completely independent of each other, giving rise to a diffusion term in constitutive equations. Typical values of ν^2 will be described in Sec. 6. The structure of steady states of system (1.1) is determined by treating $\nu^2 > 0$ as a small parameter and by applying the singular perturbation theory of [9]. This theory enables us to select steady states that appear to be appropriate for capturing the spurt phenomenon.

System (1.1) with $\nu^2 = 0$ exhibits the same behavior in steady shear as the more realistic models studied in [10–12], where the differential constitutive equations also involve normal stresses (in particular, the first normal stress difference), giving rise to a governing system of three quasi-linear parabolic-hyperbolic PDEs in place of the two in system (1.1). The dimensionless parameter α representing the ratio of Reynolds number to Deborah number is very small. The analytical study in [11–13] is based on treating the respective governing equations as singular perturbation problems with α as a singular parameter. Their approach is to determine the complete dynamics when $\alpha = 0$ and then to show that the dynamics of the full system is similar for $\alpha > 0$ sufficiently small. By contrast, our quasi-linear system (1.1) with $\nu^2 > 0$ is parabolic, and the theory of parabolic systems can be exploited to determine the global dynamics for $\alpha > 0$ sufficiently small. In particular, the existence of a global compact attractor and an inertial manifold can be established. It should be noted that the feature of mathematical models studied in [11–13] that makes their qualitative analysis (asymptotic behavior as $t \to \infty$, stability properties, etc.) particularly difficult is that the governing equations possess uncountably many isolated steady states. From this fact one can deduce that these governing systems can admit neither a compact global attractor nor a finite-dimensional inertial manifold.

The paper is organized as follows. In Sec. 2, we use general ideas from [6] to derive a non-Newtonian model of shearing motions incorporating spatial diffusion. Basic properties of the model (existence and long-time behavior of solutions, qualitative properties of steady states) are established in Sec. 3. It is shown that in the case of a generic g, the asymptotic behavior of solutions is very simple—each solution tends to some steady state and the number of steady states is finite. We also prove exponential stability of two particular steady states playing a crucial role in the explanation of spurt. In Secs. 4 and 5, spurt and hysteresis phenomena in our mathematical model are established. The phenomenon of spurt is associated with extinction of a stable steady state when the pressure gradient increases beyond a critical (bifurcation) value. The results of numerical simulations for small values of $\alpha, \nu > 0$ are presented in Sec. 6. We have performed numerical simulations of spurt and hysteresis phenomena for sample PI-3 (see [15]). Numerical results match the data observed experimentally by Vinogradov et al.

2. Non-Newtonian model of shearing motions including diffusion. In this section, we derive a mathematical model for shearing motion of a fluid leading to a system of governing equations including a diffusion term in the constitutive equation.

We consider the planar shear flow of a viscoelastic fluid in an infinite narrow strip: $x \in [-h, h]$ and $y \in (-\infty, \infty)$, with the flow directed along the y-axis. We suppose the fluid to be non-Newtonian, incompressible, and the motion to take place under isothermal conditions. We restrict ourselves to motions that are symmetric with respect to the centerline. Under our assumptions the flow variables will depend only on the transversal variable x. Hence, the velocity vector $\vec{v}$ has the form $\vec{v} = (0, v(t, x))$ with $v(t, x) = v(t, -x)$. It is easy to verify that the mass balance is then automatically satisfied. The equation governing the motion of the fluid is the balance of linear momentum

$$\varrho \left(\frac{\partial \vec{v}}{\partial t} + (\vec{v}, \nabla)\vec{v} \right) = \nabla \vec{S} \tag{2.1}$$

where ϱ is the constant fluid density and $\vec{S}$ is the total stress which can be decomposed as

$$\vec{S} = p \cdot \vec{\mathrm{Id}} + \varepsilon \cdot \vec{D} + \vec{\Sigma}. \tag{2.2}$$

Here p is the isotropic pressure of the form $p = p_0(t, x) + f \cdot y$ where f is the pressure gradient driving the flow, ε is the Newtonian viscosity, and $\vec{D}$ is the rate of deformation tensor, i.e., $\vec{D} = (\nabla \vec{v} + (\nabla \vec{v})^{\mathrm{T}})/2$. According to [6, Sec. 2] the extra stress

$$\vec{\Sigma} = \begin{pmatrix} \sigma^{xx}, & \sigma^{xy} \\ \sigma^{yx}, & \sigma^{yy} \end{pmatrix}$$

satisfies

$$\begin{aligned} \sigma^{xy} &= \sigma^{yx} = \mathscr{S}_{0s=0}^{\infty}[\Lambda_t(s)], \\ \sigma^{xx} - \sigma^{yy} &= \mathscr{S}_{1s=0}^{\infty}[\Lambda_t(s)], \\ \sigma^{xx} + \sigma^{yy} &= 0 \end{aligned} \tag{2.3}$$

where $\mathscr{S}_0, \mathscr{S}_1$ are generally nonlinear operators acting on the relative shearing history

$$\Lambda_t(s) = -\int_{t-s}^{t} v_x(\tau, x)\, d\tau. \tag{2.4}$$

Since we assume the flow to be planar, Eq. (2.1) reduces to

$$\varrho v_t = \varepsilon v_{xx} + \sigma_x + f \tag{2.5}$$

where $\sigma := \sigma^{xy}$.

We specify the operator $\mathscr{S}_0$ in such a way that it takes into account long-range molecular forces. According to [4], the latter provide the constitutive equations by a diffusion term $\nu^2 \sigma_{xx}$. The first normal stress difference determined by the operator $\mathscr{S}_1$ plays no role in our model.

Let A denote the selfadjoint closure in $L_2(0, h)$ of the operator defined on $C_B^2(0, h)$ by $Au = -u_{xx}$ for any $u \in C_B^2(0, h) := \{u \in C^2(0, h);\ u(0) = u_x(h) = 0\}$; its domain $D(A)$ is the Sobolev space $W_B^{2,2}(0, h) = \{u \in W^{2,2}(0, h);\ u(0) = u_x(h) = 0\}$. Let $\lambda, \nu > 0$ be fixed. Then the operator $-(\lambda + \nu^2 A)$ generates an analytic semigroup $\exp(-(\lambda + \nu^2 A)t)$, $t \geq 0$; (see [5, Chapter 1]).

Assume that $g: \mathfrak{R} \to \mathfrak{R}$ is an odd Lipschitz continuous function. As usual, we identify g with the Nemitsky operator $g: W^{1,2}(0, h) \to L_2(0, h)$ defined by $g(u)(x) = g(u(x))$ for a.e. $x \in [0, h]$. Due to the assumptions on g the nonlinear operator g is well defined and Lipschitz continuous.

Let $\tilde{f} \in L_2(0, h)$ be defined as

$$\tilde{f}: x \mapsto f \cdot x \quad \text{for any } x \in [0, h]. \tag{2.6}$$

We define

$$\begin{aligned} \mathscr{S}_0(\Lambda_t) = \int_0^\infty \exp(-(\lambda + \nu^2 A)s) \cdot \left[g\left(-\frac{d}{ds}\Lambda_t(s)\right) + \lambda \cdot \tilde{f} \right] ds - \tilde{f} \\ \text{for any } v \in C(\mathfrak{R}: W^{1,2}(0, h)), \ \sup_{t \in \mathfrak{R}} \|v(t)\|_{W^{1,2}} < \infty, \ \text{and } t \geq 0 \end{aligned} \tag{2.7}$$

where $\Lambda_t(s)$ is defined by Eq. (2.4), i.e., $\Lambda_t(s) = -\int_{t-s}^t v_x(\tau, x)\, d\tau$.

Clearly,

$$\mathscr{S}_0(\Lambda_t) = \int_0^\infty \exp(-(\lambda + \nu^2 A)s)[g(v_x(t - s, \cdot)) + \lambda \tilde{f}]\, ds - \tilde{f}. \tag{2.8}$$

In case $\nu = 0$, the definition of the functional $\mathscr{S}_0$ coincides with that of [6, formula (5)]. However, since the operator $\lambda + \nu^2 A$, $\nu > 0$, is a diffusion operator generating an analytic semigroup, the operator $\exp(-(\lambda + \nu^2 A)s)$, $s > 0$, smooths out solutions, i.e., $\exp(-(\lambda + \nu^2 A)s)w \in D(A)$ for any $w \in L_2(0, h)$ and $s > 0$ (see [5, Chapter 1]).

Differentiating Eq. (2.8) with respect to t and substituting $u := \sigma + \tilde{f} = \mathscr{S}_0(\Lambda_t) + \tilde{f}$, we obtain the following constitutive equation of rate type:

$$u_t + (\lambda + \nu^2 A)u = g(v_x) + \lambda \tilde{f} \tag{2.9a}$$

with boundary conditions

$$u(t, 0) = u_x(t, h) = 0 \tag{2.9b}$$

or, equivalently,

$$\sigma_t + \lambda \sigma - \nu^2 \sigma_{xx} = g(v_x) \tag{2.10a}$$

with boundary conditions

$$\sigma(t, 0) = 0, \qquad \sigma_x(t, h) = -f, \tag{2.10b}$$

respectively.

We note that $\sigma_x(t, h) = -f$ implies $v_{xx}(t, h) = 0$ which is the boundary condition appearing in the theory of multipolar fluids (see, [2, Sec. 3]). The boundary condition $u(t, 0) = 0$ $(\sigma(t, 0) = 0)$ implies that the function $u(t, \cdot)$ $(\sigma(t, \cdot))$ can be extended as an odd function to the interval $[-h, h]$ for all t. It ensures the symmetry of the flow about the centerline.

Summarizing, our model leads to the initial-boundary value problem

$$\begin{aligned} \varrho v_t &= \varepsilon v_{xx} + \sigma_x + f; \\ \sigma_t &= \nu^2 \sigma_{xx} + g(v_x) - \lambda \sigma; \\ v(0, x) = v_0(x) &\text{ and } \sigma(0, x) = \sigma_0(x) \quad \text{for a.e. } x \in [0, h]; \\ v_x(t, 0) = v(t, h) = 0, \ \sigma(t, 0) = 0, &\text{ and } \sigma_x(t, h) = -f \quad \text{for } t \geq 0. \end{aligned} \tag{2.11}$$

To facilitate the discussion, we scale the space variable x by h, times t by λ^{-1}, v by $h\lambda$, σ by $\varepsilon\lambda$, f by $\varepsilon\lambda/h$, and ν^2 by $h^2\lambda$, and replace $g(\xi)$ by $g(\lambda\xi)/\varepsilon\lambda^2$. The resulting system is

$$\begin{aligned} \alpha v_t &= v_{xx} + \sigma_x + f, \\ \sigma_t &= \nu^2 \sigma_{xx} + g(v_x) - \sigma \\ &\qquad \text{for } (t, x) \in [0, \infty] \times [0, 1] \end{aligned} \tag{2.12}$$

with boundary conditions

$$\begin{aligned} v_x(t, 0) = v(t, 1) = 0, \\ \sigma(t, 0) = 0, \qquad \sigma_x(t, 1) = -f \end{aligned} \tag{2.13}$$

and initial data

$$v(0, x) = v_0(x) \text{ and } \sigma(0, x) = \sigma_0(x) \quad \text{for a.e. } x \in [0, 1]. \tag{2.14}$$

There are two dimensionless parameters:

$$\alpha = \frac{\varrho h^2 \lambda}{\varepsilon} \quad \text{and} \quad \nu > 0.$$

According to [15] and [4], the typical values of α and ν are

$$\alpha = O(10^{-9}) \quad \text{and} \quad \nu^2 = O(10^{-4}).$$

Hence, we may treat α and ν as small parameters.

3. Existence of solutions, asymptotic behavior, steady-state solutions and their stability. In this section, we study the problem of existence of solutions, their long-time behavior, and some qualitative properties of steady states of the system (2.12). Using the abstract theory developed in [5] we establish local and global solvability. For g real analytic we furthermore prove that the asymptotic behavior of the solutions is simple—each trajectory approaches some steady state and the number of steady state solutions is finite. To single out the appropriate stationary solutions, we apply the results of the theory of singularly perturbed boundary value problems of [9].

3.1. *Existence of solutions.* In terms of the variables v and u the initial boundary value problem (2.12) takes the form

$$\begin{aligned} \alpha v_t &= v_{xx} + u_x, \\ u_t &= \nu^2 u_{xx} - u + g(v_x) + fx, \end{aligned} \tag{3.1}$$

$$\begin{aligned} v_x(t, 0) = v(t, 1) = 0 \text{ and } u(t, 0) = u_x(t, 1) = 0 \quad \text{for } t \geq 0, \\ v(0, x) = v_0(x) \text{ and } u(0, x) = u_0(x) \quad \text{for } x \in [0, 1]. \end{aligned}$$

To facilitate the discussion, let

$$S = v_x + u = v_x + \sigma + \tilde{f}. \tag{3.2}$$

Obviously,

$$\alpha S_t = S_{xx} + \alpha u_t. \tag{3.3}$$

In terms of S and u, the system (3.1) takes the form

$$\begin{aligned} \alpha S_t &= S_{xx} + \alpha\nu^2 u_{xx} + \alpha(g(S-u) + fx - u), \\ u_t &= \nu^2 u_{xx} - u + g(S-u) + fx \end{aligned} \tag{3.4}$$

with boundary conditions

$$u(t, 0) = u_x(t, 1) = 0, \qquad S(t, 0) = S_x(t, 1) = 0$$

and initial data

$$S(0, x) = S_0(x) = v_{0x}(x) + u_0(x), \text{ and } u(0, x) = u_0(x) \quad \text{for } x \in [0, 1]. \tag{3.5}$$

Throughout this paper we will assume that α and ν are small parameters. The pressure gradient f is assumed to be positive. The function $h(u) := u + g(u)$ is assumed to be C^2 with a single loop as shown in Fig. 1.

More precisely, we make the following hypotheses:

(i) $g: \mathfrak{R} \to \mathfrak{R}$ is an odd C^2 function with bounded first and second derivatives satisfying $g(u)u \geq 0$ for any $u \in \mathfrak{R}$;

(ii) there exist constants $0 < c_1 < c_2$ such that

$$\begin{aligned} &h'(u) = 1 + g'(u) > 0, \ h'' < 0 \quad \text{on } [0, c_1), \\ &h'(u) = 1 + g'(u) < 0 \quad \text{on } (c_1, c_2), \\ &h'(u) = 1 + g'(u) > 0, \ h'' > 0 \quad \text{on } (c_2, \infty). \end{aligned} \tag{W}$$

Under assumptions (W), there exists a $\gamma_0 > 0$ such that

$$\int_{\min h^{-1}(\gamma_0)}^{\max h^{-1}(\gamma_0)} (h(u) - \gamma_0)\, du = 0.$$

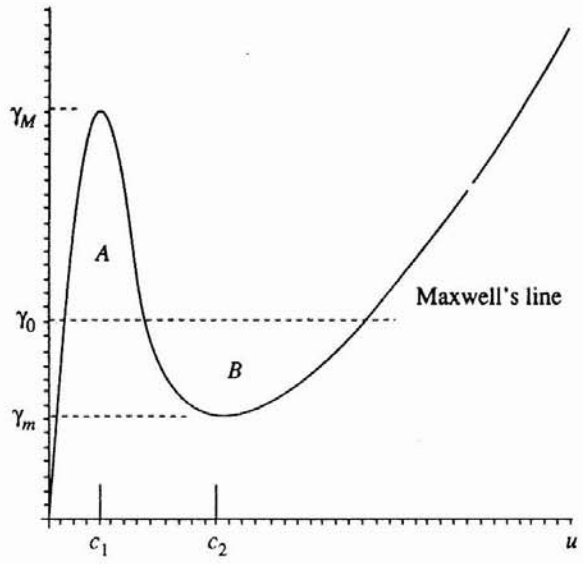


FIG. 1. van der Walls type curve

The last integral condition is commonly known as *Maxwell's equal area rule* (the area A equals B). In Fig. 1 the line $u = \gamma_0$ is called *Maxwell's line.* We also note that the function $h(u) = u + g(u)$ satisfying (W) is sometimes called *van der Walls type curve.*

In what follows, we let X denote the real Hilbert space $L_2(0, 1)$ with norm $\|\cdot\|$ and inner product $(\cdot, \cdot)$. Recall that the operator A defined in the previous section is sectorial and positive in X with domain $D(A) = \{w \in W^{2,2}(0, 1);\ w(0) = w_x(1) = 0\}$. Hence, fractional powers of A can be defined. Let X^γ, $\gamma \geq 0$, be the Hilbert space consisting of the domain $D(A^\gamma)$ endowed with the graph norm

$$\|w\|_\gamma = \|A^\gamma w\| \text{ for any } w \in X^\gamma = D(A^\gamma). \tag{3.6}$$

The operator A has a compact resolvent $A^{-1}: X \to X$.

Now one can treat the governing equations (3.4), (3.5) as abstract differential equations in the Hilbert space

$$\mathscr{X} = X \times X. \tag{3.7}$$

To do so, we let $\Phi = \begin{bmatrix} S \\ u \end{bmatrix}$. The system (3.4) then becomes

$$\frac{d}{dt}\Phi + L\Phi = F(\Phi), \qquad \Phi(0) = \Phi_0 = \begin{bmatrix} S_0 \\ u_0 \end{bmatrix} \tag{3.8}$$

where the linear operator L is defined by

$$L\begin{bmatrix} S \\ u \end{bmatrix} := \begin{bmatrix} A(\frac{1}{\alpha}S + \nu^2 u) \\ \nu^2 A u \end{bmatrix} = \begin{pmatrix} \frac{1}{\alpha}A & \nu^2 A \\ 0 & \nu^2 A \end{pmatrix}\begin{bmatrix} S \\ u \end{bmatrix} \tag{3.9}$$

on its domain $D(L) = D(A) \times D(A)$. The nonlinearity F is given by

$$F\left(\begin{bmatrix} S \\ u \end{bmatrix}\right) = \begin{bmatrix} g(S-u) - u + fx \\ g(S-u) - u + fx \end{bmatrix}. \tag{3.10}$$

It is routine to verify that $L: D(L) \subset \mathscr{X} \to \mathscr{X}$ is a sectorial operator generating an analytic semigroup $\exp(-Lt)$, $t \geq 0$. Since A^{-1} is compact, it is easy to show that L has a compact resolvent $L^{-1}: \mathscr{X} \to \mathscr{X}$. The fractional power $L^{1/2}$ is then easily computed as

$$\begin{pmatrix} \frac{1}{\sqrt{\alpha}}A^{1/2} & \frac{\sqrt{\alpha}\nu^2}{1+\nu\sqrt{\alpha}}A^{1/2} \\ 0 & \nu A^{1/2} \end{pmatrix}$$

and $D(L^{1/2}) = D(A^{1/2}) \times D(A^{1/2})$. Hence there is an equivalent norm in $\mathscr{X}^{1/2}$ such that

$$\mathscr{X}^{1/2} \cong X^{1/2} \times X^{1/2}, \tag{3.11}$$

and it can easily be verified that

$$X^{1/2} = \{w \in W^{1,2}(0, 1);\ w(0) = 0\}. \tag{3.12}$$

Since we have assumed that the first and second derivative of g are bounded, the nonlinearity F is a C^1 mapping from $\mathscr{X}^{1/2}$ into $\mathscr{X}$.

Now we can apply the general theory of abstract parabolic equations [5]. According to [5, Theorems 3.3.3, 3.3.4, 3.4.1, and 3.5.2], for any initial condition $\Phi_0 \in \mathscr{Z}^{1/2}$ the abstract equation (3.8) has a unique solution $\Phi(t)$ defined on $[0, \infty)$ by the property

$$\Phi \in C_{\text{loc}}([0, \infty), \mathscr{Z}^{1/2}) \cap C^1_{\text{loc}}((0, \infty), \mathscr{Z}^{1/2}),$$
$$\Phi(t) \in D(L) \text{ for } t > 0 \quad \text{and } \Phi(0) = \Phi_0.$$

Hence, Eq. (3.8) defines a C^1-semidynamical system $(T(t),\ t \geq 0)$ in $\mathscr{Z}^{1/2}$ defined by

$$T(t)\Phi_0 = \Phi(t, \Phi_0) \quad \text{for any } t \geq 0$$

where $\Phi(t, \Phi_0)$ is the solution of Eq. (3.8) with $\Phi(0) = \Phi_0 \in \mathscr{Z}^{1/2}$.

3.2. *Asymptotic behavior of solutions.* We now turn our attention to the asymptotic behavior of solutions of Eq. (3.8). First, we will study the set of steady states, i.e., stationary solutions of Eq. (3.8) which we denote by $\mathscr{E}$. Clearly,

$$\mathscr{E} = \left\{ \begin{bmatrix} 0 \\ \overline{u} \end{bmatrix};\ \overline{u} \in D(A) \text{ is a solution of } \nu^2 A\overline{u} = -\overline{u} + g(-\overline{u}) + fx \right\}. \tag{3.13}$$

In fact, $[\begin{smallmatrix} 0 \\ \overline{u} \end{smallmatrix}] \in \mathscr{E}$ iff

$$\overline{u} \in C^4(0, 1), \qquad \nu^2 \overline{u}_{xx} + \overline{u} + g(\overline{u}) - fx, \qquad \overline{u}(0) = \overline{u}_x(1) = 0. \tag{3.14}$$

Here we have used the assumption that g is an odd C^2 function.

The system (3.8) admits a global Lyapunov function $V: \mathscr{Z}^{1/2} \to \mathfrak{R}$ defined by

$$V\left(\begin{bmatrix} S \\ u \end{bmatrix}\right) = \frac{1}{2}\left\{\frac{1}{\alpha}\|S\|^2_{1/2} + \nu^2\|S - u\|^2_{1/2} + \|S - u\|^2 + J(S - u)\right\}$$

where

$$J(w) = 2\int_0^1 \int_0^{w(x)} (g(s) + fx)\, ds\, dx. \tag{3.15}$$

Indeed, a simple calculation shows that for any solution $[\begin{smallmatrix} S(t) \\ u(t) \end{smallmatrix}]$ the following formula holds:

$$\frac{d}{dt} V\left(\begin{bmatrix} S(t) \\ u(t) \end{bmatrix}\right) + \frac{1}{\alpha}\|S(t)\|^2_{1/2} + \frac{1 + \alpha\nu^2}{\alpha^2}\|S(t)\|^2_1 = 0 \quad \text{for any } t > 0. \tag{3.16}$$

Due to the assumption $g(u)u \geq 0$ for any $u \in \mathfrak{R}$ it follows that the functional V is bounded from below. From Eqs. (3.14), (3.16) it follows that the real-valued function $t \mapsto V([\begin{smallmatrix} S(t) \\ u(t) \end{smallmatrix}])$, $t \geq 0$, is strictly decreasing unless $[\begin{smallmatrix} S(t) \\ u(t) \end{smallmatrix}] = [\begin{smallmatrix} 0 \\ \overline{u} \end{smallmatrix}] \in \mathscr{E}$ is a steady-state solution of Eq. (3.8). Then a standard argument (see, e.g., [16, Theorem 4.1]) enables us to conclude that the omega-limit set

$$\Omega(\Phi_0) := \{\Phi \in \mathscr{Z}^{1/2},\ \text{there exists } t_n \to \infty \text{ such that } T(t_n)\Phi_0 \to \Phi\}$$

satisfies

$$\Omega(\Phi_0) \subseteq \mathscr{E}, \tag{3.17}$$

for any $\Phi_0 \in \mathscr{X}^{1/2}$. Since the operator L has a compact resolvent L^{-1}, it follows from [5, Theorems 3.3.6 and 4.3.3] and Eq. (3.17) that

$$\lim_{t\to\infty} \operatorname{dist}(T(t)\Phi_0, \mathscr{E}) = 0, \tag{3.18}$$

where $\operatorname{dist}(\Phi, \mathscr{E}) = \inf(\|\Phi - \Psi\|_{\mathscr{X}^{1/2}}, \ \Phi \in \mathscr{E})$. In the following simple proposition, we obtain bounds on steady states, and we show for g real analytic that the number of possible steady states is finite.

PROPOSITION 3.1. Let $u_0 \geq c_2$ be such that $h(u_0) \geq f$. Then $0 \leq u(x) \leq u_0$ for any solution $u(x)$ of Eq. (3.14). Moreover, there exists a constant $M = M(g, f) > 0$ such that

$$\nu \sup_{x\in[0,1]} |u_x(x)| + \sup_{x\in[0,1]} |u(x)| \leq M.$$

If g is real analytic, then the number of solutions of Eq. (3.14) is finite.

Proof. Let u be an arbitrary solution of Eq. (3.14). Since $h(u) := u + g(u)$ is nondecreasing on $[u_0, \infty)$ and $h(u_0) \geq f$, it follows that $u(x) \geq u_0$ implies $\nu^2 u_{xx}(x) = h(u(x)) - fx \geq h(u_0) - fx \geq f(1-x)$. Thus the function $u(x)$ is strictly convex whenever $u(x) \geq u_0$. Since $u(0) = 0$, if $u(x_0) > u_0$ for some $x_0 \in (0, 1]$, then there exists $x_1 \in (0, 1)$ such that $u(x_1) = u_0$, $u(x) > u_0$, and $u_x(x) > 0$ on $(x_1, 1)$. This means that u cannot satisfy $u_x(1) = 0$. Hence, $u(x) \leq u_0$ for every $x \in [0, 1]$ and $\nu > 0$. The inequality $0 \leq u(x)$ can be obtained in a similar way. The estimates for $u(x)$ and $\nu u_x(x)$ follow from the well-known interpolation inequality

$$\nu \sup_{x\in[0,1]} |u_x(x)| \leq 2\left(\sup_{x\in[0,1]} |u(x)| + \nu^2 \sup_{x\in[0,1]} |u_{xx}(x)|\right)$$

for any $u \in C^2([0, 1])$ and $\nu > 0$.

Now we assume that g is real analytic. We fix a $\nu > 0$ and define the map $\mu \mapsto \phi(\mu)$ as $\phi(\mu) = u_x^\mu(1)$ where $u^\mu(x)$ is the solution of the initial-value problem $\nu^2 u_{xx} = u + g(u) - fx$, $u^\mu(0) = 0$, $u_x^\mu(0) = \mu$. Since g is Lipschitz continuous and analytic, the function $\phi(\mu)$ is well defined and analytic on $\mathfrak{R}$. Furthermore, $\phi(\mu) = 0$ if and only if $u^\mu(x)$ is a solution of the BVP (3.14). Suppose to the contrary, the existence of infinitely many solutions of the BVP (3.14). Then the set $\{\mu \in [-M/\nu, M/\nu];\ \phi(\mu) = 0\}$ must have an accumulation point. Because of analyticity of ϕ, we have $\phi \equiv 0$ on $\mathfrak{R}$. Hence, there is a solution $u^\mu(x)$ of the BVP (3.14) for $\mu > M/\nu$ which is inconsistent with $u_x^\mu(0) = \mu$. □

The omega-limit set $\Omega(\Phi_0)$ is connected [5, Theorem 4.3.3]. Thus, by Eq. (3.17), $\Omega(\Phi_0)$ is a singleton whenever $\mathscr{E}$ is finite. We have thus established the following.

THEOREM 3.2. Assume the hypotheses (W). Then, for any initial condition $\Phi_0 \in \mathscr{X}^{1/2}$, the evolution problem (3.8) has the unique solution $\Phi = \Phi(t, \Phi_0)$, $t \geq 0$, its omega-limit set $\Omega(\Phi_0)$ being contained in the set of steady-state solutions $\mathscr{E}$. If, in addition, g is real analytic, then each trajectory tends to a single steady state.

3.3. *Steady-state solutions.* We now examine steady-state solutions of Eq. (3.8). Recall that $\begin{bmatrix}\overline{S}\\ \overline{u}\end{bmatrix}$ is a steady state if and only if $\overline{S} \equiv 0$ and $\overline{u} \in C^4(0, 1)$ is a solution

of the BVP

$$\begin{aligned} \nu^2 u_{xx} &= u + g(u) - fx, \\ u(0) &= u_x(1) = 0. \end{aligned} \tag{3.19}$$

The steady-state velocity profile $\overline{v}$ is then calculated as $\overline{v}(x) = \int_x^1 \overline{u}(\xi)\, d\xi$. Since ν is assumed to be small, the problem (3.19) can be viewed as a singular perturbation of the reduced problem

$$0 = u + g(u) - fx. \tag{3.20}$$

From now on, we assume

$$f \in [f_{\min}, f_{\max}],$$

where $0 < f_{\min} < \gamma_m$ and $\gamma_M < f_{\max} < \infty$. From Fig. 1 it is clear that the problem (3.20) has a unique C^1 solution $u = \phi_1(x)$, $x \in [0, 1]$, whenever $f \in [f_{\min}, \gamma_m)$. When $f \in [\gamma_m, f_{\max}]$ there exist C^1 functions $\phi_i(x)$ defined on two overlapping intervals I_i contained in $[0, 1]$, where $0 \in I_1$, $1 \in I_2$, $i = 1, 2$, and such that $h(\phi_i(x)) - fx = 0$, $x \in I_i$, and $\phi_2(x) > \phi_1(x)$ on $I_1 \cap I_2$. Hence, there also exist discontinuous solutions of (3.20). Indeed, any function $u = u(x)$ where $u = \phi_1(x)$ on $[0, 1]\backslash I_2$, $u(x) \in \{\phi_1(x), \phi_2(x)\}$ on $I_1 \cap I_2$ and $u = \phi_2(x)$ on $[0, 1]\backslash I_1$ is the solution of (3.20); the number of discontinuities of u is unlimited. Inevitably, each solution of (3.20) is discontinuous whenever $f \in (\gamma_M, f_{\max}]$. In the case $f \in (\gamma_0, f_{\max}]$ and ν small we expect the existence of a solution of (3.19) having an abrupt transition at some interior point $x_0 \in (0, 1)$. When ϕ_1 is defined on the whole interval $[0, 1]$ we also expect that (3.19) has a solution that is close to ϕ_1 on $[0, 1]$ for ν small.

To make the above discussion precise, we employ general results of singularly perturbed equations due to Lin [9]. To this end, let us consider (3.19) as the equivalent 2×2 system

$$\begin{aligned} \nu u_x &= w, \\ \nu w_x &= u + g(u) - fx, \\ u(0) &= w(1) = 0. \end{aligned} \tag{3.21}$$

In case $f \in [f_{\min}, \gamma_M)$ the piecewise continuous function

$$\overline{U}_\nu^1 = \begin{cases} (0, 0), & x \in [0, \nu^{1/2}), \\ (\phi_1(x), 0), & x \in [\nu^{1/2}, 1 - \nu^{1/2}), \\ (\phi_1(1), 0), & x \in [1 - \nu^{1/2}, 1] \end{cases} \tag{3.22}$$

is a formal approximation of the system (3.21) in the sense of [9, Theorem 2.1]. When $f \in (\gamma_0, f_{\max}]$ (γ_0 is determined by Maxwell's equal area rule), there is another formal approximation of system (3.21) given by

$$\overline{U}_\nu^2 = \begin{cases} (0, 0), & x \in [0, \nu^{1/2}); \\ (\phi_1(x), 0), & x \in [\nu^{1/2}, x_0 - \nu^{1/2}]; \\ (z(\frac{x - x_0}{\nu}),\ z'(\frac{x - x_0}{\nu})), & x \in (x_0 - \nu^{1/2}, x_0 + \nu^{1/2}); \\ (\phi_2(x), 0), & x \in [x_0 + \nu^{1/2}, 1 - \nu^{1/2}); \\ (\phi_2(1), 0), & x \in [1 - \nu^{1/2}, 1]. \end{cases} \tag{3.23}$$

Here $x_0 \in (0, 1)$ is determined by $fx_0 = \gamma_0$ and $z = z(\tau)$ is the heteroclinic solution of the second-order autonomous ODE

$$z'' = z + g(z) - \gamma_0 \tag{3.24}$$

such that $\lim_{\tau\to-\infty} z(\tau) = \phi_1(x_0)$, $\lim_{\tau\to\infty} z(\tau) = \phi_2(x_0)$, $z > 0$, and $z' > 0$. The existence of such a solution follows (by phase-plane analysis) from the fact that (due to the hypothesis (W)) $\phi_1(x_0)$ and $\phi_2(x_0)$ lie on the same level curve of an integral for the system (3.21). We note that $\phi_1(x_0) = \min h^{-1}(\gamma_0)$, $\phi_2(x_0) = \max h^{-1}(\gamma_0)$ for any $f \in [\gamma_0, f_{\max}]$, and hence the solution z does not depend on f.

It is now easy to verify that the formal approximations $\overline{U}_\nu^{(1)}$ and $\overline{U}_\nu^{(2)}$ satisfy the hypotheses (H1)–(H3) of [9]. We omit this detail. Then the main result of [9] adapted to the BVP (3.19) reads

THEOREM 3.3 [9, Theorem 2.2]. Let $\overline{U}_\nu$ be a formal approximation of (3.19) given by (3.22) or (3.23). Then there exists $\nu_0 > 0$ and $\delta_0 > 0$ such that for $0 < \nu \le \nu_0$ there exists a unique true solution $u = u_\nu(x)$ of system (3.19) with $r := \sup_{x\in[0,1]} |U_\nu(x) - \overline{U}(x)| \le \delta_0$, where $U_\nu(x) = (u(x), \nu u_x(x))$. The remainder r is of order $O(\nu^{1/2})$ when $\nu \to 0^+$.

REMARK 3.4. Theorem 2.2 of [9], however, does not specify the explicit dependence of the remainder r on the coefficients of Eq. (3.19). The decay of the remainder r may depend on the parameter f. Nevertheless, for any fixed $\eta > 0$ small enough, using the implicit function theorem and following the lines of the proof of [9, Theorems 2.2, 4.3, and 4.4], one can show that the remainder $r = r(\nu, f)$ for the formal approximation $\overline{U}_\nu^{(1)}$ ($\overline{U}_\nu^{(2)}$) is $O(\nu^{1/2})$ uniformly with respect to $f \in [f_{\min}, \gamma_M - \eta]$ and $f \in [\gamma_0 + \eta, f_{\max}]$, respectively, when $\nu \to 0^+$.

For $f \in [f_{\min}, \gamma_M)$, Theorem 3.3 asserts the existence of a true solution $u_\nu^{(1)}$ of Eq. (3.19) approximating the given formal approximation $\overline{U}_\nu^1$. We have

$$u_\nu^{(1)}(x) \xrightarrow{\text{unif}} \phi_1(x) \text{ and } v_\nu^{(1)}(x) \xrightarrow{\text{unif}} \int_x^1 \phi_1(\xi)\,d\xi \quad \text{for any } x \in [0, 1] \text{ as } \nu \to 0^+. \tag{3.25}$$

Again, by Theorem 3.3, for any $f \in (\gamma_0, f_{\max}]$, there exists a solution $u_\nu^{(2)}$ of Eq. (3.19) such that

$$\lim_{\nu\to0^+} u_\nu^{(2)}(x) = \phi_1(x) \quad \text{for any } x \in [0, x_0),$$

$$\lim_{\nu\to0^+} u_\nu^{(2)}(x) = \phi_2(x) \quad \text{for any } x \in (x_0, 1]. \tag{3.26}$$

Hence, for small $\nu > 0$ the solution $u_\nu^{(2)}$ has a graph as in Fig. 2 (see p. 412).

By the Lebesgue dominated convergence theorem we have the uniform convergence

$$v_\nu^{(2)} \xrightarrow{\text{unif}} v_0^{(2)} \equiv \begin{cases} \int_x^1 \phi_2(\xi)\,d\xi, & x \in [x_0, 1]; \\ \int_x^{x_0} \phi_1(\xi)\,d\xi + \int_{x_0}^1 \phi_2(\xi)\,d\xi, & x \in [0, x_0] \end{cases}$$

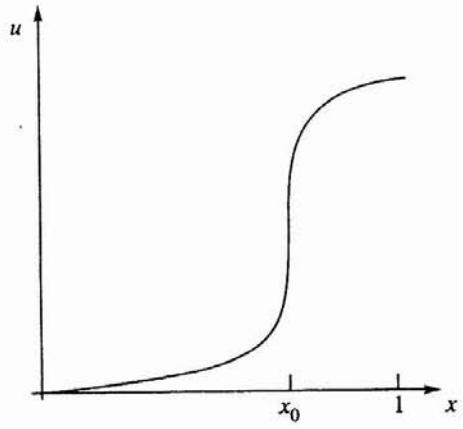


FIG. 2

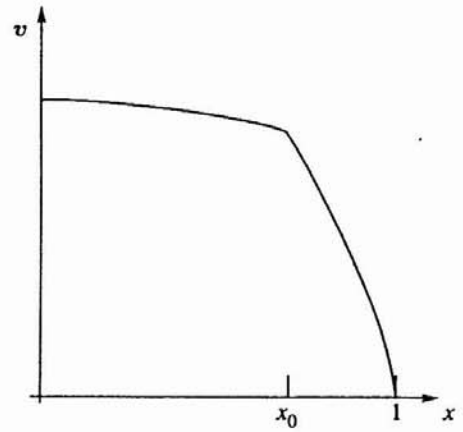


FIG. 3

when $\nu \to 0^+$. Hence, the family $(v_\nu^{(2)})_{\nu>0}$ converges uniformly to the velocity profile $v_0^{(2)}$ with a kink located at x_0 as shown in Fig. 3.

It is now clear that given a pressure gradient $f \in (\gamma_0, \gamma_M)$, for any ν sufficiently small there exist at least two solutions $u_\nu^{(1)}$, $u_\nu^{(2)}$ of Eq. (3.19) satisfying Eqs. (3.25) and (3.26), respectively.

Integrating the velocity $\overline{v}$ with respect to x yields the steady-state flow rate per cross section

$$Q = 2\int_0^1 \overline{v}(x)\,dx. \tag{3.27}$$

Denote by Q_ν^i the volumetric flow rate corresponding to the velocity $v_\nu^{(i)}$ given by Eqs. (3.25) and (3.26), respectively. Clearly, for any $\eta > 0$ there is $d = d(g, \eta) > 0$ such that

$$Q_\nu^{(2)} - Q_\nu^{(1)} \geq d \quad \text{for any } f \in [\gamma_0 + \eta, \gamma_M) \text{ and } \nu > 0 \text{ sufficiently small.} \tag{3.28}$$

We conclude this section by discussing the stability of steady states. We first show that linearized stability of a solution $\bar{u}$ of system (3.19) extends to that of the steady-state solution $\begin{bmatrix} 0 \\ \bar{u} \end{bmatrix}$ of Eq. (3.8).

LEMMA 3.5. Let $0 < \alpha < 1/\sup_{u \in \mathfrak{R}} |g'(u)|$. A steady-state solution $\begin{bmatrix} 0 \\ \bar{u} \end{bmatrix}$ of Eq. (3.8) is exponentially asymptotically stable with respect to small perturbations of initial data in the phase space $\mathscr{X}^{1/2} = X^{1/2} \times X^{1/2}$, provided the principal eigenvalue μ_0 of the linearized Sturm-Liouville problem $B_1[u] = \nu^2 u_{xx} - u - g'(-\bar{u}(x))u = \mu u$, $u(0) = u_x(1) = 0$ is negative.

Using Lemma 3.5 we are able to prove the theorem below establishing stability of the solutions $\begin{bmatrix} 0 \\ u_\nu^i \end{bmatrix}$, $i = 1, 2$, as well as their uniqueness for certain parameter values. The details of the proofs of Lemma 3.5 and Theorem 3.6 are given in the appendix.

THEOREM 3.6. Assume that $0 < \alpha < 1/\sup_{u \in \mathfrak{R}} |g'(u)|$ and g satisfies the hypotheses (W).

(a) If $f \in [f_{\min}, \gamma_M)$ and $\nu > 0$ is sufficiently small, then the principal eigenvalue μ_0 of the linearized Sturm-Liouville problem $B_1[u] = \mu u$ at $u_\nu^{(2)}$ is negative. Consequently, the steady-state solution $\begin{bmatrix} 0 \\ u_\nu^{(1)} \end{bmatrix}$ of Eq. (3.8) is exponentially asymptotically stable with respect to small perturbations of initial data in the phase space $\mathscr{X}^{1/2} = X^{1/2} \times X^{1/2}$.

(b) If $f \in (\gamma_0, f_{\max}]$ and $\nu > 0$ is sufficiently small, then the principal eigenvalue μ_0 of the linearized Sturm-Liouville problem $B_1[u] = \mu u$ at $u_\nu^{(1)}$ is negative. Consequently, the steady-state solution $\begin{bmatrix} 0 \\ u_\nu^{(2)} \end{bmatrix}$ of Eq. (3.8) is exponentially asymptotically stable with respect to small perturbations of initial data in the phase space $\mathscr{X}^{1/2} = X^{1/2} \times X^{1/2}$.

(c) There exists a unique steady-state solution of Eq. (3.8) whenever $f \in [f_{\min}, \gamma_m)$ or $f \in (\gamma_M, f_{\max}]$ and $\nu > 0$ is sufficiently small.

4. Spurt. Having developed the mathematical background we are in position to explain the occurrence of spurt for a fluid governed by the system of equations (3.8).

Suppose that we are loading the pressure gradient quasi-statically from $f_{\min}$ to $f_{\max}$ allowing the system to settle down to its equilibrium state at each step.

Since $v_\nu^{(1)} = v_\nu^{(1)}(f)$ depends continuously on f, the volumetric flow rate $Q_\nu^{(1)} = Q_\nu^{(1)}(f)$ of the steady-state velocity $v_\nu^{(1)} = v_\nu^{(1)}(f)$ for $f < \gamma_M$ forms a continuous curve. At each step of the "loading-stabilization" procedure, the volumetric flow rate corresponding to the velocity $v(T)$ is close to $Q_\nu^{(1)} = Q_\nu^{(1)}(f)$ when T is large enough.

The situation changes dramatically when the pressure gradient f passes γ_M. For $f > \gamma_M$ the solution has no other possibility than to settle down to the unique steady-state solution

$$\begin{bmatrix} 0 \\ u_\nu^{(2)}(\cdot, f) \end{bmatrix}$$

of system (3.8) which is globally asymptotically stable by Theorem 3.6. Hence, by Eq. (3.28), this small change of the pressure gradient causes a jump of size $d > 0$ in the volumetric flow rate as shown in Fig. 4. This jump is equal to the area between the two equilibrium solutions $v_\nu^{(1)}$ and $v_\nu^{(2)}$ (see Fig. 4).

For f varying in the interval $(\gamma_M, f_{\max}]$, the "loading-stabilization" can be repeated. The corresponding volumetric flow rates are close to the continuous curve $f \mapsto Q_\nu^{(2)}(f)$ of the steady-state volumetric flow rates in Fig. 5.

Let us note that earlier models that did not include the diffusion terms in their constitutive relations also captured the spurt phenomenon [10–12]. For $f > \gamma_M$ the principal difference between our explanation of spurt and that of papers mentioned is: the change in volumetric flow rate as f passes through the critical value γ_M on loading is much more drastic in our model than the earlier ones; here the "kink" develops at the point $0 < \gamma_0/\gamma_M < 1$ very suddenly and then moves slowly with a definite speed toward the centerline. In [10, 11], the kink develops at the wall; for $f > \gamma_M$, the layer position is $x^* = \gamma_M/f$. The phenomenon of latency that occurs on loading described in [10, 11] is not discussed here.

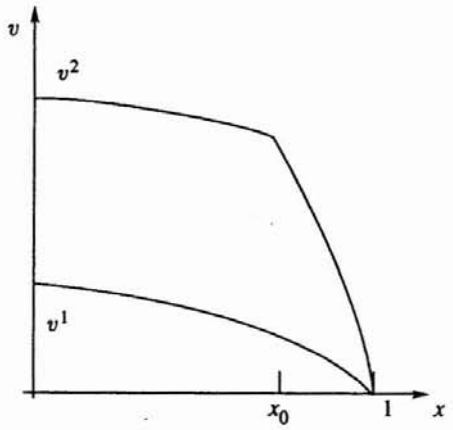


FIG. 4

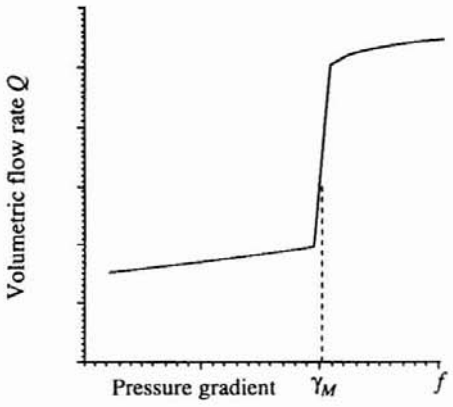


FIG. 5. Spurt

5. Hysteresis. We now consider the loading-unloading cyclic process. The behavior of the volumetric flow rate during the loading period has been described in the previous section. Recall that the volumetric flow rate increased rapidly when the pressure gradient passed the value γ_M. Now let us unload the pressure gradient starting from $f = f_{\max}$. By convention, as long as f stays larger than γ_0, the solution still settles down on

$$\begin{bmatrix} v_\nu^{(2)}(\cdot, f) \\ u_\nu^{(2)}(\cdot, f) \end{bmatrix}.$$

On the other hand, for any $f < \gamma_m$ there exists the unique solution

$$\begin{bmatrix} v_\nu^{(1)}(\cdot, f) \\ u_\nu^{(1)}(\cdot, f) \end{bmatrix}.$$

Therefore, the solution

$$\begin{bmatrix} v_\nu^{(2)}(\cdot, f) \\ u_\nu^{(2)}(\cdot, f) \end{bmatrix}$$

ceases to exist at some critical value near γ_0. Figure 6 shows two branches of the bifurcation diagram corresponding to the stable steady states

$$\begin{bmatrix} v_\nu^{(i)}(\cdot, f) \\ u_\nu^{(i)}(\cdot, f) \end{bmatrix}, \qquad i = 1, 2.$$

By Eq. (3.28), $Q_\nu^{(2)}(f) - Q_\nu^{(1)}(f) \geq d(\eta) > 0$ for any $f \subset [\gamma_0 + \eta, \gamma_M)$ where $\eta > 0$ is fixed. Hence, there is a hysteresis loop as shown in Fig. 7 (see p. 416).

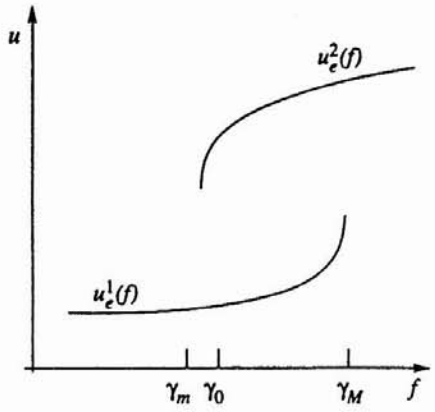


FIG. 6

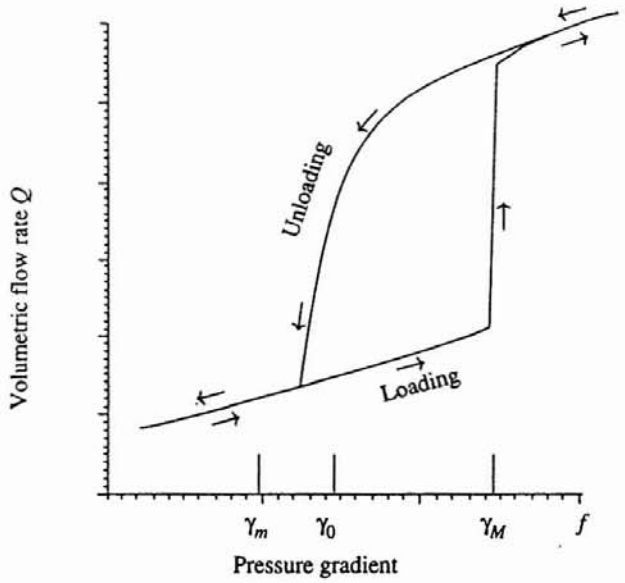


FIG. 7. Hysteresis

6. Numerical simulations. In this section we present some numerical results exhibiting spurt and hysteresis. Recall that our model leads to the system of governing equations

$$\begin{aligned} \varrho v_t &= \varepsilon v_{xx} + \sigma_x + f; \\ \sigma_t &= \nu^2 \sigma_{xx} + g(v_x) - \lambda\sigma \end{aligned} \qquad (6.1)$$

$$\text{for } (t, x) \in [0, \infty] \times [0, r_{\text{cap}}]$$

with boundary conditions

$$v_x(t, 0) = v(t, r_{\text{cap}}) = 0, \quad \sigma(t, 0) = 0, \qquad \sigma_x(t, r_{\text{cap}}) = -f$$

and initial data

$$v(0, x) = v_0(x) \text{ and } \sigma(0, x) = \sigma_0(x) \quad \text{for a.e. } x \in [0, r_{\text{cap}}]. \qquad (6.2)$$

We will consider an analytic function g of a particular form

$$g(u) = \mu \frac{u}{1 + (1 - a^2)u^2/\lambda^2} \qquad (6.3)$$

where $\mu > 0$ is the elastic modulus, a is the dimensionless slip parameter, and λ is the relaxation time of the polymer. The particular choice of the function g is taken from [11, Sec. 3].

First, we determine the magnitude of the coefficient $\nu > 0$ in Eqs. (6.1). Following [4]

$$\nu^2 \approx \frac{k \cdot \theta}{2\xi} \qquad (6.4)$$

where θ is the absolute temperature, k is the Boltzmann constant, and ξ is the hydrodynamic resistance of one dumb-bell bead (assumed to be constant). If we take typical values of $\theta \approx 10^2$K, $\xi \approx 10^{-9}$kg s^{-1} and recall that $k \approx 10^{-23}$J K^{-1}, we

obtain $\nu^2 \approx 10^{-12}\text{m}^2\ \text{s}^{-1}$. In our numerical simulations we have chosen the fixed value

$$\nu^2 = 4 \times 10^{-12}\text{m}^2\ \text{s}^{-1}. \tag{6.5}$$

We next turn to the Vinogradov et al. rheological data. In all experiments, the radius of the capillary was

$$r_{\text{cap}} = 0.48 \times 10^{-3}\text{m}.$$

The elastic modulus μ and the density ϱ have been taken constant for all samples and equal to

$$\mu = 6 \times 10^4\text{Pa}, \qquad \varrho = 10^3\text{kg m}^{-3}, \tag{6.6}$$

respectively.

Numerical experiments were performed for the polyisoprene PI-3 which was the first sample for which spurt was observed [15, Fig. 3b]. According to [15] and [8, p. 323] we have

$$\lambda = 0.1\ \text{s}^{-1}, \qquad \varepsilon = 0.01484\frac{\mu}{\lambda} = 8.9 \times 10^3\text{Pa s}^{-1} \qquad a = 0.98. \tag{6.7}$$

We see that the constants $\alpha = \varrho r_{\text{cap}}^2\lambda/\varepsilon = 2.58 \times 10^{-9}$ and $\nu^2/r_{\text{cap}}^2\lambda = 10^{-4}$ introduced in Sec. 2 can be treated as small parameters. It is easy to verify that the real analytic function

$$h(u) = \lambda u + \frac{\mu}{\varepsilon} \cdot \frac{u}{1 + (1 - a^2)u^2/(\varepsilon^2\lambda^2)}$$

is of van der Walls type (see the hypothesis (W)).

As our first numerical experiment, we simulated spurt. In S. I. units, we choose

$$f_{\min} = 9.3 \times 10^7\text{kg m}^{-2}\ \text{s}^{-2}, \qquad f_{\max} = 51.2 \times 10^7\text{kg m}^{-2}\ \text{s}^{-2},$$
$$\Delta f = 1.8 \times 10^7\text{kg m}^{-2}\ \text{s}^{-2}.$$

The startup initial condition (for $f = f_{\min}$) was chosen to be $(v_0, u_0) = (0, 0)$. At each loading step, the solutions were followed for a sufficiently long time $T_{\max} = 150$ sec to allow them to settle down. Since $\alpha > 0$ was very small, we could use the Crank-Nicholson implicit time-space discretization scheme. The spatial mesh contained a total of 40 nodes. The time step was chosen as $\Delta t = 0.005$ sec.

Figure 8 (see p. 418) shows the results obtained (Fig. 8(a)) and compares them with Vinogradov et al.'s experimental data (Fig. 8(b), the flow curve for PI-3 is labeled by 3). Following [15] c-g-s units are employed and axes are in the logarithmic scale. The nominal shear stress τ is defined by $\tau = r_{\text{cap}}f$ (see [8, Eq. (48)]). Since we have considered a planar flow instead of a capillary flow the corresponding definition of a volumetric flow rate is

$$Q = \frac{3}{r_{\text{cap}}^2}\int_0^{r_{\text{cap}}} v(x)\,dx$$

(see [8, Eq. (47)]).

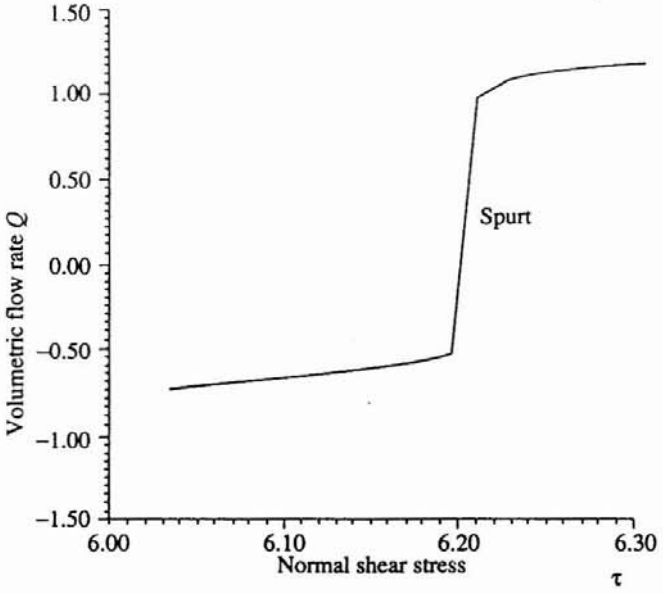


FIG. 8(a). The spurt phenomenon for the sample PI-3.

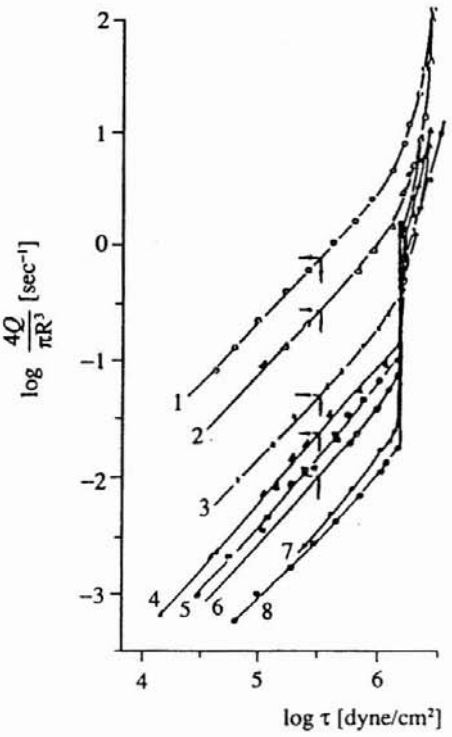


FIG. 8(b).

Finally, we have performed numerical simulations of a loading-unloading cycle. The hysteresis loop under the cyclic load is displayed in Fig. 9.

Figure 10 shows the steady, kinked velocity profile for the spurt value of the nominal shear stress $\tau = 1.61 \times 10^6 \text{dyne cm}^{-2}$ $(\log \tau = 6.21)$.

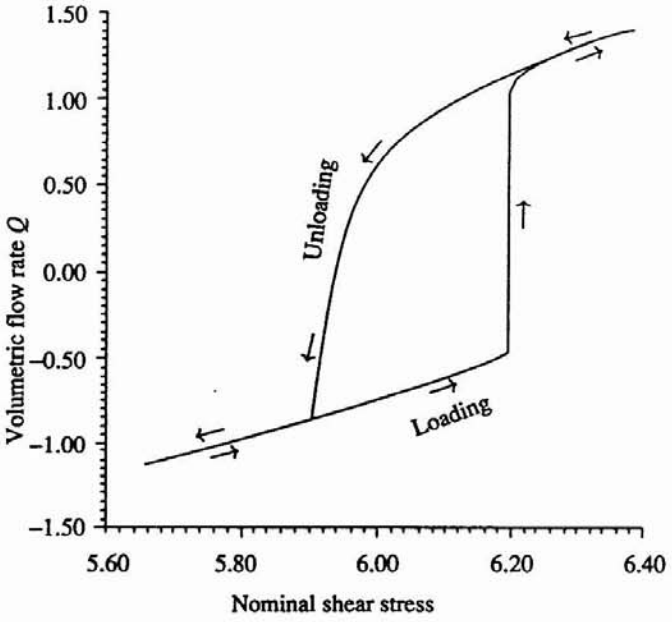


FIG. 9. The hysteresis loop under cyclic load

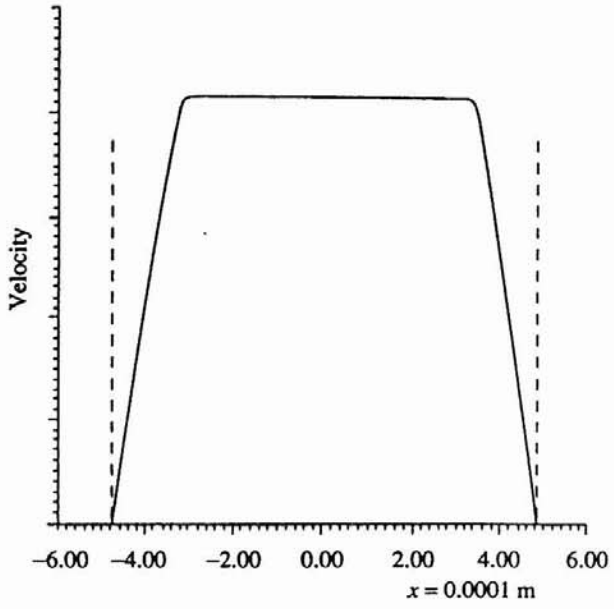


FIG. 10. The velocity profile at the critical value of pressure

7. Discussion. We have proposed a modification of the mathematical model of shearing motions leading to a system of governing equations including a diffusion term $\nu^2\sigma_{xx}$ in the constitutive equation. In addition, we have described the asymptotic behavior of solutions which is simple in typical situations—each solution tends to some steady state and the number of steady states is finite.

The diffusion term makes the system of governing equations parabolic. As a consequence of the resulting parabolic smoothing effect the system will admit a finite-dimensional inertial manifold as well as a compact global attractor. In a subsequent paper we will study singular limits when $\alpha = \varrho r_{\text{cap}}^2 \lambda/\varepsilon$ tends to zero.

Acknowledgments. The authors are thankful to J. A. Nohel and A. Tzavaras for introducing them to the subject and helpful discussions.

Appendix.

Proof of Lemma 3.5. Let $\left[\begin{smallmatrix}0\\ \overline{u}\end{smallmatrix}\right]$ be an arbitrary steady state solution of Eq. (3.8). The linearization of Eq. (3.8) at $\left[\begin{smallmatrix}0\\ \overline{u}\end{smallmatrix}\right]$ has the form

$$\frac{d}{dt}\begin{bmatrix} S \\ u \end{bmatrix} = B\begin{bmatrix} S \\ u \end{bmatrix}$$

where the linear operator B is given by

$$B\begin{bmatrix} S \\ u \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha}S_{xx} + \nu^2 u_{xx} - u + g'(-\overline{u}(x))(S-u) \\ \nu^2 u_{xx} - u + g'(-\overline{u}(x))(S-u) \end{bmatrix}, \tag{A.1}$$

its domain being $D(B) = \{\left[\begin{smallmatrix}S\\ u\end{smallmatrix}\right],\ S, u \in W^{2,2}(0,1);\ S(0) = S_x(1) = u(0) = u_x(1) = 0\} \subset L_2(0,1) \times L_2(0,1)$. Denote by B_1 the Sturm-Liouville operator

$$B_1[u] = \nu^2 u_{xx} - u - g'(-\overline{u}(x))u \tag{A.2}$$

on its domain $D(B_1) = \{w \in W^{2,2}(0,1);\ w(0) = w_x(1) = 0\} \subset L_2(0,1)$.

Assume that the principal eigenvalue μ_0 of the linear problem $B_1[u] = \mu u$, $u \in D(B_1)$ is negative. Since B_1 is a selfadjoint Sturm-Liouville operator, we have

$$\frac{(B_1[u], u)}{\|u\|^2} \le \mu_0 < 0 \tag{A.3}$$

for any $u \in D(B_1)$, $u \neq 0$. Moreover, B_1 is invertible and $B_1^{-1}: L_2 \to L_2$ is compact. Hence, the operator B is also invertible and

$$B^{-1}\begin{bmatrix} \phi \\ \psi \end{bmatrix} = \begin{bmatrix} \alpha A^{-1}(\psi - \phi) \\ B_1^{-1}(\psi - \alpha g'(-\overline{u}(\cdot))A^{-1}(\psi - \phi)) \end{bmatrix}$$

where the linear operator A was defined in Sec. 2. Since, by Eq. (3.6), $A^{-1}: L_2 \to L_2$ is compact, $B^{-1}: \mathscr{X} \to \mathscr{X}$ is compact as well. Therefore, the spectrum $\sigma(B)$ consists of eigenvalues.

We will show that $\operatorname{Re}\lambda < 0$ for any $\lambda \in \sigma(B) = \sigma_p(B)$. Suppose to the contrary that there exists an eigenvalue $\lambda \in \sigma(B)$ such that $\operatorname{Re}\lambda \geq 0$. Let $\begin{bmatrix} S \\ u \end{bmatrix}$ denote the eigenvector of the linear problem

$$B\begin{bmatrix} S \\ u \end{bmatrix} = \lambda \begin{bmatrix} S \\ u \end{bmatrix}. \tag{A.4}$$

Subtracting the equations for S and u we obtain $\frac{1}{\alpha} S_{xx} = \lambda(S-u)$. Thus,

$$S_x(x) = -\alpha\lambda \int_x^1 (S-u)(\xi)\, d\xi. \tag{A.5}$$

Taking the inner product of (A.5) with $-\int_x^1 (S-u)(\xi)\, d\xi$ we obtain

$$-\|S-u\|^2 - (u, S-u) = \alpha\lambda \left\| \int_x^1 (S-u)(\xi)\, d\xi \right\|^2.$$

Since $\operatorname{Re}\lambda \geq 0$, we have $\|S-u\|^2 \leq -\operatorname{Re}(u, S-u) \leq \|u\|\,\|S-u\|$ and hence,

$$\|S-u\| \leq \|u\|. \tag{A.6}$$

From (A.5) we have $S(x) = -\alpha\lambda \int_0^x \int_r^1 (S-u)(\xi)\, d\xi\, dr$. Thus $S = \alpha\lambda\, J(S-u)$ where $J: L_2 \to L_2$ is a linear bounded operator with $\|J\| \leq 1$ Therefore, u satisfies the equation

$$B_1[u] + \alpha\lambda g'(-\overline{u}(\cdot))J(S-u) = \lambda u. \tag{A.7}$$

Take the inner product of (A.7) with u to obtain

$$(B_1[u], u) = \lambda(\|u\|^2 - \alpha(g'(-\overline{u}(\cdot))J(S-u), u)).$$

Since B_1 is selfadjoint, we have $\operatorname{Im}(\lambda - \alpha\lambda(g'(-\overline{u}(\cdot))J(S-u), u)/\|u\|^2) = 0$ and

$$\mu_0 \geq \frac{(B_1[u], u)}{\|u\|^2} = \lambda\left(1 - \alpha\frac{(g'(-\overline{u}(\cdot))J(S-u), u)}{\|u\|^2}\right).$$

According to (A.6) we have

$$\alpha\left|\frac{(g'(-\overline{u}(\cdot))J(S-u), u)}{\|u\|^2}\right| \leq \alpha \sup_{s\in\mathfrak{R}} |g'(s)| \frac{\|J(S-u)\|\,\|u\|}{\|u\|^2} \leq \alpha \sup_{s\in\mathfrak{R}} |g'(s)| < 1$$

because $\|J\| \leq 1$. Therefore,

$$\mu_0 \geq \lambda\left(1 - \alpha\frac{(g'(-\overline{u}(\cdot))J(S-u), u)}{\|u\|^2}\right) \geq 0,$$

a contradiction. Hence, $\operatorname{Re}\lambda < 0$ for any $\lambda \in \sigma(B)$. By [5, Theorem 5.1.1], the steady-state solution $\begin{bmatrix} 0 \\ \overline{u} \end{bmatrix}$ of Eq. (3.8) is exponentially asymptotically stable with respect to small perturbations of initial data in the phase space $\mathscr{X}^{1/2} = X^{1/2} \times X^{1/2}$. □

Proof of Theorem 3.6. (a) For any $u \in D(B_1)$, $u \neq 0$, we have

$$\begin{aligned} \frac{(B_1[u], u)}{\|u\|^2} &= \frac{1}{\|u\|^2}\left(-\nu^2 \int_0^1 u_x^2(x)\, dx - \int_0^1 h'(u_\nu^{(1)}(x))u^2(x)\, dx\right) \\ &\leq -\frac{1}{\|u\|^2}\int_0^1 h'(u_\nu^{(1)}(x))u^2(x)\, dx. \end{aligned} \tag{A.8}$$

We have $h'(\phi_1(x)) > 0$ for $x \in [0, 1]$. Therefore, $h'(u_\nu^{(1)}(x)) > 0$ for any $x \in [0, 1]$ and ν small. Hence, the principal eigenvalue μ_0 of B_1 satisfies

$$\mu_0 = \sup_{u \in D(B_1),\ u \neq 0} \frac{(B_1[u], u)}{\|u\|^2} < 0. \tag{A.9}$$

(b) Let us now consider the solution $u_\nu^{(2)}$ of Eq. (3.19) having an abrupt transition at the point $x_0 = \gamma_0/f \in (0, 1)$.

First we prove that $u_\nu^{(2)}$ is increasing on $[0, 1)$. The curve $h(u) - fx = 0$ splits the first quadrant into two parts (Fig. A.1).

The function $u_\nu^{(2)}$ is convex or concave at x depending on whether the point $(x, u_\nu^{(2)}(x))$ belongs to the left-hand or to the right-hand component labeled by $+, -$, respectively. According to Theorem 3.3 we have

$$\sup\{|u_\nu^{(2)}(x) - \phi_1(x)|,\ x \in [0, x_0 - \nu^{1/2}]\} = O(\nu^{1/2}),$$

$$\sup\{|u_\nu^{(2)}(x) - \phi_2(x)|,\ x \in [x_0 + \nu^{1/2}, 1]\} = O(\nu^{1/2})$$

as $\nu \to 0^+$. Since $u_\nu^{(2)}$ is a solution of Eq. (3.19) and $0 \leq u_\nu^{(2)}$ (by Proposition 3.1), we have $\frac{d}{dx}u_\nu^{(2)}(0) > 0$. Indeed, $\frac{d}{dx}u_\nu^{(2)}(0) \leq 0$ would imply

$$\frac{d^3}{dx^3}u_\nu^{(2)}(0) = \frac{1}{\nu^2}\left(h'(u_\nu^{(2)}(0))\frac{d}{dx}u_\nu^{(2)}(0) - f\right) < 0.$$

Since $u_\nu^{(2)}(0) = \frac{d^2}{dx^2}u_\nu^{(2)}(0) = 0$, we have $u_\nu^{(2)}(x) < 0$ for some $x > 0$, a contradiction. By an obvious indirect argument, one can show that $\frac{d}{dx}u_\nu^{(2)}(x)$ cannot

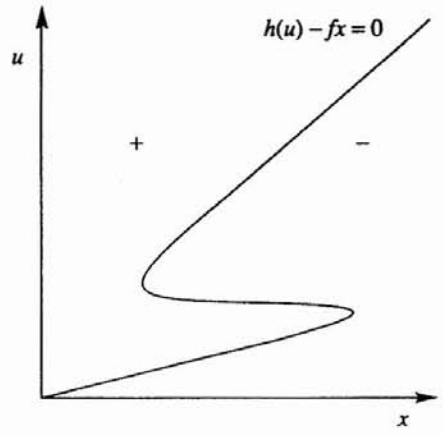


FIG. A.1.

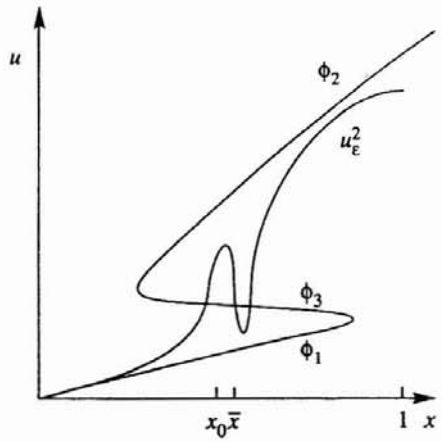


FIG. A.2.

become negative in $[0, x_0-\nu^{1/2}]\cup[x_0+\nu^{1/2}, 1]$. To prove that $\frac{d}{dx}u_\nu^{(2)}$ is positive in $(x_0-\nu^{1/2}, x_0+\nu^{1/2})$ suppose the contrary. Since $u_\nu^{(2)}$ is convex in + and concave in −, this is possible only if there exists an $\bar{x}\in(x_0-\nu^{1/2}, x_0+\nu^{1/2})$ such that $\frac{d}{dx}u_\nu^{(2)}(\bar{x})<0$ and $u_\nu^{(2)}(\bar{x})=\phi_3(\bar{x})$, ϕ_3 being the middle branch solution of $h(u)-fx=0$ is shown in Fig. A.2.

Let us introduce the "fast-time" variable $\tau=(x-x_0)/\nu$ for $x\in(x_0-\nu^{1/2}, x_0+\nu^{1/2})$ and put $u(\tau)=u_\nu^{(2)}(x_0+\nu\tau)$. Then $\frac{d}{d\tau}u(\tau)=\nu\frac{d}{dx}u_\nu^{(2)}(x_0+\nu\tau)$. According to Theorem 3.3 we have

$$\sup_{\tau\in(-\nu^{-1/2},\nu^{-1/2})}\left|\frac{d}{d\tau}(u(\tau)-z(\tau))\right|=O(\nu^{1/2})\quad\text{as }\nu\to0^+,$$

z being the heteroclinic solution of the problem (3.24). Since $\bar{x}-x_0=O(\nu^{1/2})$, we have $|\phi_3(\bar{x})-\phi_3(x_0)|=O(\nu^{1/2})$ as $\nu\to0^+$. Therefore, $\frac{d}{dx}u_\nu^{(2)}(\bar{x})=\nu\frac{d}{d\tau}u((\bar{x}-x_0)/\nu)$ must have the same sign as $\frac{d}{d\tau}z((\bar{x}-x_0)/\nu)$ for any ν small. Hence $\frac{d}{dx}u_\nu^{(2)}(\bar{x})>0$, a contradiction.

Knowing that for any $f\in(\gamma_0, f_{\max}]$, $u_\nu^{(2)}$ is increasing in $[0,1)$ for ν small we return to the linearized eigenvalue problem $B_1[u]=\mu u$ where $B_1[u]=\nu^2u_{xx}-h'(u_\nu^{(2)}(x))u$, $u(0)=u_x(1)=0$. First we prove the following useful lemma.

LEMMA A. Assume $f\in[f_{\min}, f_{\max}]$. Let $\bar{u}$ be any nondecreasing solution of (3.19) such that $|h(\bar{u}(1))-f|<(1-a)f$ and $h'(\bar{u}(x))\geq0$ on $[a,1]$ for some $a\in(0,1)$. Then the principal eigenvalue μ_0 of the linear operator $B_1[w]=\nu^2w_{xx}-h'(\bar{u}(x))w$, $w\in D(B_1)$, is negative.

Proof. Denote $\phi(x)=\frac{d}{dx}\bar{u}(x)$. Then ϕ satisfies

$$\nu^2\phi_{xx}-h'(\bar{u}(x))\phi=-f,\qquad\phi_x(0)=\phi(1)=0,\tag{A.10}$$

and $\phi > 0$ on $[0, 1)$. Let w be a solution of

$$B_1[w] = \nu^2 w_{xx} - h'(\overline{u}(x))w = \mu_0 w, \qquad w(0) = w_x(1) = 0 \tag{A.11}$$

corresponding to the principal eigenvalue μ_0 of B_1. Since (A.11) is a Sturm-Liouville problem, there exists w satisfying (A.11) such that $w > 0$ on $(0, 1)$ and $\int_0^1 w(x)\,dx = 1$. If we multiply (A.11) by ϕ and integrate over $[0, 1]$, we obtain

$$\begin{aligned}\mu_0 \int_0^1 w(x)\phi(x)\,dx &= \nu^2(w_x\phi - w\phi_x)|_0^1 - f\int_0^1 w(x)\,dx \quad [\text{because } w_x(0)\phi(0) \geq 0] \\ &\leq -w(1)(h(\overline{u}(1) - f) - f \leq w(1)|h(\overline{u}(1)) - f| - f.\end{aligned} \tag{A.12}$$

Now suppose to the contrary that $\mu_0 \geq 0$. Since $w > 0$ on $(0, 1)$, $w_x(1) = 0$, we have $\nu^2 w_{xx} = h'(\overline{u}(x))w + \mu_0 w \geq 0$ on $[a, 1]$. Hence, $w(x) \geq w(1)$ on $[a, 1]$ and, consequently,

$$1 = \int_0^1 w(x)\,dx \geq \int_a^1 w(x)\,dx \geq (1 - a)w(1).$$

From (A.12) we obtain

$$\mu_0 \int_0^1 w(x)\phi(x)\,dx < 0.$$

Since $w \geq 0$, $\phi \geq 0$, we have $\mu_0 < 0$, a contradiction. □

Now it is easy to complete the proof of part (b). We fix an $a > x_0$. Then, by Theorem 3.3, $\sup\{|u_\nu^{(2)}(x) - \phi_2(x)|,\ x \in [a, 1]\} = O(\nu^{1/2})$ as $\nu \to 0^+$. Therefore, $|h(u_\nu^{(2)}(1)) - f| < (1 - a)f$ and $h'(u_\nu^{(2)}(x)) > 0$ on $[a, 1]$ for any $\nu > 0$ sufficiently small. Lemma A completes the proof.

Note that, for certain singularly perturbed problems, an asymptotic estimate of the form $\mu_0(\nu) = O(\nu)$ as $\nu \to 0^+$ is proved in [1].

(c) Our next goal is to prove uniqueness of solutions of (3.19) for $f \in [f_{\min}, \gamma_m) \cup (\gamma_M, f_{\max}]$ and ν small. Let us consider the case $f \in (\gamma_M, f_{\max}]$. First, we show linearized stability of an arbitrary nondecreasing solution $\overline{u}$ of (3.19). By Lemma A it is sufficient to prove that $|h(\overline{u}(1)) - f| < (1 - a)f$ and $h'(\overline{u}(x)) \geq 0$ on $[a, 1]$ for some $a \in (0, 1)$. To this end, we recall first that according to Proposition 3.1 there exists an $M > 0$ such that

$$\nu \sup_{x\in[0,1]} |\overline{u}_x(x)| + \sup_{x\in[0,1]} |\overline{u}(x)| \leq M \tag{A.13}$$

for any solution $\overline{u}$ of (3.19) and $\nu > 0$.

Let $\overline{u}$ be a nondecreasing solution for (3.19). Let $1 > \tilde{a} > \gamma_M/f$. Then for any $x \in [\tilde{a}, 1]$ we have $fx > \gamma_M$; so $\overline{u}$ is concave on $[\tilde{a}, 1]$. Thus, by (A.13)

$$0 \leq \overline{u}_x(x) \leq \int_{\tilde{a}}^x \overline{u}_x(\xi)\,d\xi \cdot \frac{1}{x - \tilde{a}} \leq \frac{4M}{1 - \tilde{a}} \tag{A.14}$$

for any $x \in [a, 1]$ where $a = (\tilde{a} + 1)/2$. Therefore, there exists a constant $M_1 > 0$ such that

$$0 \leq fx - h(\overline{u}(x)) \leq f\xi - h(\overline{u}(\xi)) + M_1(\xi - x) \tag{A.15}$$

for any $\xi, x \in [a, 1]$, $x \le \xi$. Thus, by (A.14) and (A.15)

$$\begin{aligned} 0 &\le \nu^{1/2}(fx - h(\overline{u}(x))) \\ &\le \int_x^{x+\nu^{1/2}} (f\xi - h(\overline{u}(\xi)) + M_1(\xi - x))\, d\xi \\ &= -\nu^2 \int_x^{x+\nu^{1/2}} \overline{u}_{xx}(\xi)\, d\xi + \frac{M_1\nu}{2} \le \left(2M + \frac{M_1}{2}\right)\nu =: M_2\nu. \end{aligned}$$

Hence $|fx - h(\overline{u}(x))| \le M_2\nu^{1/2}$ for any $x \in [a, 1]$, $\nu > 0$, and any nondecreasing solution $\overline{u}$ of (3.19).

For $\nu \le ((fa - \gamma_M)/M_2)^2$ we have

$$h(\overline{u}(x)) \ge fx - |fx - h(\overline{u}(x))| \ge fa - |fa - \gamma_M| = \gamma_M \quad \text{for any } x \in [a, 1].$$

Since $h(u) \le \gamma_M$ for $u \le c_2$ (see Fig. 1), we have $\overline{u}(x) \ge c_2$ on $[a, 1]$, hence $h'(\overline{u}(x)) \ge 0$ for $x \in [a, 1]$. By Lemma A, the principal eigenvalue μ_0 of the problem $B_1[w] = \nu^2 w_{xx} - h'(\overline{u}(x))w = \mu w$, $w \in D(B_1)$, is negative.

Now, consider the parabolic equation

$$u_\tau = \nu^2 u_{xx} - h(u) + fx,$$

$$u(\tau, 0) = u_x(\tau, 1) = 0, \qquad \tau \ge 0, \ u(0, x) = u_0(x), \ x \in [0, 1].$$

This equation generates a gradient-like semidynamical system $\mathscr{S}(\tau)$, $\tau \ge 0$, in the Hilbert space $X^{1/2} = \{u \in W^{1,2}(0, 1), \ u(0) = 0\}$ defined by $\mathscr{S}(\tau)u_0 = u(\tau, \cdot)$, where $u(0, \cdot) = u_0(\cdot)$ (see [5, Chapter 4]). The set $\mathscr{K} = \{u \in X^{1/2}, \ u_x(x) \ge 0$, a.e. on $[0, 1]\}$ is a closed convex cone in $X^{1/2}$. Moreover, $\mathscr{K}$ is invariant under $\mathscr{S}$, i.e.,

$$u(\tau, \cdot) \in \mathscr{K} \quad \text{whenever } u(0, \cdot) \in \mathscr{K} \text{ for any } \tau \ge 0.$$

Indeed, the function

$$w(\tau, x) = \begin{cases} -u_x(\tau, x), & x \in [0, 1], \ \tau \ge 0; \\ -u_x(\tau, -x), & x \in [-1, 0], \ \tau \ge 0, \end{cases}$$

is the solution of the scalar parabolic equation

$$w_\tau = \nu^2 w_{xx} - h'(u(x))w - f,$$
$$w(\tau, -1) = w(\tau, 1) = 0.$$

Therefore, $w(\tau, x) \le 0$ whenever $w(0, x) \le 0$ by the Maximum Principle (see [14]). Hence, $\mathscr{S}$ is a semidynamical system on the complete metric space $\mathscr{K}$ with the topology induced by $X^{1/2}$.

To complete the proof we argue similarly as in [1, Theorem 4]. Since $\mathscr{K}$ is invariant, it is the union of (disjoint) attraction domains of the nondecreasing stationary solutions of (A.14). Because those solutions are asymptotically stable, these attraction domains are open in $\mathscr{K}$. Since the set $\mathscr{K}$ is connected, it cannot be a union of two nonempty disjoint open sets; hence, $u_\nu^{(2)}$ is the unique stationary solution in $\mathscr{K}$.

Now, let $\overline{u}$ be an arbitrary solution of (3.19) (not necessarily nondecreasing). By Proposition 3.1, $\overline{u}$ is bounded and $\overline{u} \geq 0$. Then there exist $\overline{u}^-$, $\overline{u}^+ \in \mathscr{K} \cap D(A)$ such that $\overline{u}^-(x) \leq \overline{u}(x) \leq \overline{u}^+(x)$, $x \in [0, 1]$. With regard to the Maximum Principle [14, Chapter 3, Theorem 3] we obtain $\mathscr{S}(\tau)\overline{u}^-(x) \leq \mathscr{S}(\tau)\overline{u}(x) \leq \mathscr{S}(\tau)\overline{u}^+(x)$ for any $\tau \geq 0$ and $x \in [0, 1]$. Since $\mathscr{S}(\tau)\overline{u}^\pm \in \mathscr{K}$, for any $\tau \geq 0$, we have $\mathscr{S}(\tau)\overline{u}^\pm \to u_\nu^{(2)}$ as $\tau \to \infty$. Thus, $\overline{u} = u_\nu^{(2)}$.

Hence, the solution $u_\nu^{(2)}$ is unique, provided ν is small and $f \in (\gamma_M, f_{\max}]$. The proof of uniqueness of solutions of (3.19) for $f \in [f_{\min}, \gamma_m)$ is similar. □

References

[1] S. Angenent, J. Mallet-Paret, and L. A. Pelletier, *Stable transition layers in a semilinear boundary value problems*, J. Differential Equations **67**, 212–242 (1987)

[2] H. Bellout, F. Bloom, and J. Nečas, *Phenomenological behavior of multipolar viscous fluids*, Quart. Appl. Math. **50**, 559–583 (1992)

[3] A. V. Bhave, R. C. Armstrong, and R. A. Brown, *Kinetic theory and rheology of dilute, nonhomogeneous polymer solutions*, J. Chem. Phys. **87**, 3024–3025 (1991)

[4] A. W. El-Kareh and G. L. Leal, *Existence of solutions for all Deborah numbers for a non-Newtonian model modified to include diffusion*, J. Non-Newtonian Fluid Mech. **33**, 257–287 (1989)

[5] D. Henry, *Geometric theory of semilinear parabolic equations*, Lecture Notes in Math., vol. 840, Springer-Verlag, New York, 1981

[6] J. K. Hunter and M. Slemrod, *Viscoelastic fluid flow exhibiting hysteretic phase changes*, Phys. Fluids **26**, 2345–2351 (1983)

[7] R. Kolkka and G. Ierley, *Phase space analysis of the spurt phenomenon fluid for the Giesekus viscoelastic fluid model*, J. Non-Newtonian Fluid Mech. **33**, 305–323 (1989)

[8] R. Kolkka, D. Malkus, D. Hansen, G. Ierley, and R. Worthing, *Spurt phenomena of the Johnson-Sagelman fluid and related models*, J. Non-Newtonian Fluid Mech. **29**, 303–325 (1988)

[9] Xiao-Biao Lin, *Shadowing lemma and singularly perturbed boundary value problems*, SIAM J. Appl. Math. **49**, 26–54 (1989)

[10] D. S. Malkus, J. A. Nohel, and B. J. Plohr, *Analysis of new phenomena in shear flow of non-Newtonian fluids*, SIAM J. Appl. Math. **51**, 899–929 (1991)

[11] D. S. Malkus, J. A. Nohel, and B. J. Plohr, *Dynamics of shear flow of a non-Newtonian fluid*, J. Comp. Phys. **87**, 464–487 (1990)

[12] J. A. Nohel and R. L. Pego, *Nonlinear stability and asymptotic behavior of shearing motions of a non-Newtonian fluid*, SIAM J. Math. Anal. **24**, 911–942 (1993)

[13] J. A. Nohel, R. L. Pego, and A. E. Tzavaras, *Stability of discontinuous steady states in shearing motion of a non-Newtonian fluid*, Proc. Roy. Soc. Edinburgh Sect. A **115**, 39–59 (1990)

[14] M. Protter and H. F. Weinberger, *Maximum principles in differential equations*, Springer-Verlag, New York, 1984

[15] G. Vinogradov, A. Malkin, Yu. Yanovskii, E. Borisenkova, B. Yarlykov, and G. Berezhnaya, *Viscoelastic properties and flow of narrow distribution polybutadienes and polyisoprenes*, J. Polymer Sci. Part A-2 **10**, 1061–1084 (1972)

[16] G. F. Webb, *Existence and asymptotic behavior for a strongly damped nonlinear wave equation*, Canad. Math. J. **32**, 631–643 (1980)

P. Brunovský, A. Erdélyi, H.-O. Walther

On a model of a currency exchange rate—local stability and periodic solutions

J. Dynam. Differential Equations 16(2) (2004), 393–432.

Journal of Dynamics and Differential Equations, Vol. 16, No. 2, April 2004 (© 2004)

On a Model of a Currency Exchange Rate – Local Stability and Periodic Solutions*

Pavol Brunovský[1], Alexander Erdélyi[2,3] and Hans-Otto Walther[4,5]

Received December 31, 2003

A delay differential equation is presented which models how the behavior of traders influences the short time price movements of an asset. Sensitivity to price changes is measured by a parameter a. There is a single equilibrium solution, which is non-hyperbolic for all $a>0$. We prove that for $a<1$ the equilibrium is asymptotically stable, and that for $a>1$ a 2-dimensional global center-unstable manifold connects the equilibrium to a periodic orbit. Its birth at $a=1$ is not of Hopf type and seems part of a Takens–Bogdanov scenario.

KEY WORDS: Delay differential equation; periodic solution; center manifold reduction; Takens–Bogdanov singularity.

1. INTRODUCTION

In this paper we study the equation

$$\dot{x}(t)=a(x(t)-x(t-1))-|x(t)|x(t) \tag{1.1}$$

with $a>0$ which is obtained from the equation

$$\dot{x}(t)=a(x(t)-x(t-1))-b|x(t)|x(t) \tag{1.2}$$

---

* Dedicated to Professor Shui-Nee Chow on the occasion of his 60th birthday.

[1] Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia.

[2] Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia.

[3] Currently CERGE-EI, Prague, Czech Republic.

[4] Mathematisches Institut, Universität Gießen, Gießen, Germany. E-mail: Hans-Otto.Walther@math.uni-giessen.de

[5] To whom correspondence should be addressed.

with $a > 0, b > 0$ by the normalization $x \mapsto bx$. By Eq. (1.2) we model short-time fluctuations of an asset the price of which is freely determined by supply and demand and reacts quickly to their movement. For definiteness we will speak about the exchange rate (i.e., price of a foreign currency in a domestic reference one) under a floating rate policy, although the model applies equally well to other kinds of assets (like e.g., share or commodity prices).

Economic theory provides several models for the dependence of exchange rate on macroeconomic parameters, like e.g., the purchasing power parity (PPP) index. Those parameters change slowly and, therefore, cannot be accountable for fluctuations of time constant of days one can observe. The purpose of the model is to single out possible source of those fluctuations – the psychology of agents trading the foreign currency.

To this end we suppose that there is a "natural" macroeconomically justified equilibrium exchange rate which we assume to be constant within our time range. We assume that agents have a certain perception but not precise knowledge of this rate and denote by x the deviation from it.

In order to make profit from their trading, agents attempt to forecast the movement of the exchange rate in the future. As a source for their forecast they employ its movement in the immediate past and a feeling about its position with respect to the equilibrium rate. On one hand, they expect the rate to raise if it did so in the past. So, a raise of the exchange rate leads to increasing demand and thus an increase of the price. This mechanism is expressed by the first term on the left-hand side of Eq. (1.2). On the other hand, once the rate moves (for definiteness, raises) farther from its equilibrium more and more agents expect that this trend will eventually turn back. This leads to increasing supply and hence a decrease of the rate. We model this mechanism by the second, quadratic term, assuming that both the number of agents and the amount supplied by an individual representative agent increase.

For a more thorough discussion of the model the reader is referred to [2]. Let us note that in [4] the discrete time version of the same equation is developed, although by different reasoning.

The following asymptotic behavior of solutions is indicated by numerical experiments: For $a \leqslant 1$ the equilibrium is globally asymptotically stable whereas for $a > 1$ it is unstable. Solutions tend to 0 eventually monotonically for $a < 1$, for $a = 1$ they oscillate around 0 with the distance of their zeros tending to ∞ eventually monotonically. For $a > 1$ there is a globally stable periodic orbit the period of which tends to ∞ for $a \to 1$.

The economic implications of these dynamics are discussed in detail in [2]. Here we just briefly mention its principal message: In case the sensitivity (in economic terms, elasticity) of the agents to the increase/decrease

of the exchange rate exceeds the threshold value $a = 1$, the equilibrium loses its stability and permanent fluctuations occur. In other words, under a certain configuration of parameters no other reason except of the psychology of agents is needed for permanent fluctuations of the exchange rate.

The principal goal of this paper was to establish, at least partially, the numerically observed dynamics rigorously. However, the equation turned out to be interesting mathematically in itself. Its only equilibrium $x(t) \equiv 0$ is non-hyperbolic for all values of the parameters, and for the critical value $a = 1$ it has a double zero eigenvalue linearization of Bogdanov–Takens type [13]. When a is increased and passes the critical value 1 the dimension of the center unstable manifold (which is the center manifold for $a <$ 1) jumps from 1 to 2.

Mathematical intuition is somewhat in conflict with the economic one, the latter turning out to be more adequate: Whereas linearization suggests monotone solutions, by economic intuition the conflicting mechanisms should lead to fluctuations. As another surprise, the (decisive) action of the seemingly stabilizing nonlinear term of the equation turns out to be repelling in the center manifold for $a > 1$.

The paper succeeds in two ways: To a considerable extent we can establish the observed local dynamics at zero, and we prove the existence of a periodic orbit for $a > 1$. Whereas the former employs advanced but well known methods (center manifold reduction in particular), the proof of existence of a periodic orbit contains new ingredients, compared to earlier work [16, 17, 19]. Among others, it shows how planar curves associated with solutions as in the Poincaré–Bendixson type results [3, 10, 11, 14, 15] are precisely related to the solution curves in the infinite-dimensional state space. Details on this are given in the final Section 8.

We end this section by introducing some notation and other preliminaries. Section 2 provides simple facts about boundedness, compactness, and injectivity of solution operators. Section 3 contains elementary results on oscillatory behavior of solutions. Section 4 is devoted to the linearization at 0 and prepares the local and global studies in the sequel. The local center manifold reduction stability analysis follows in Section 5. Section 6 deals with a global center-unstable manifold in case $a > 1$. In Section 7 it is shown that the boundary of the global center-unstable manifold is a periodic orbit. In the concluding Section 8 we explain what is new in the proof of existence of a periodic solution, exhibit the relation to Poincaré–Bendixson type results, and discuss open problems of mathematical as well as economic interest.

Notation, preliminaries. Let B be a Banach space. For any real number $r > 0$, the open ball of radius r centered at $0 \in B$ is denoted by B_r.

For a subset $M \subset B$, the closure, boundary, and interior are denoted by $\bar{M}, \partial M, \overset{\circ}{M}$, respectively. A *compact* map from a subset of a Banach space into a Banach space maps bounded sets into sets with compact closure. Spectra of linear operators from a subspace of a Banach space B over $\mathbb{R}$ into B are defined via complexifications.

C denotes the Banach space of all continuous real functions on the interval $[-1, 0]$, with the norm given by $\|\phi\| = \max_{-1 \leqslant t \leqslant 0} |\phi(t)|$. C^1 denotes the Banach space of all continuously differentiable real functions on $[-1, 0]$, with the norm given by $\|\phi\|_1 = \|\phi\| + \|\dot{\phi}\|$. The embedding $J : C^1 \ni \phi \mapsto \phi \in C$ is compact. For any continuous function $f : \mathbb{R} \to \mathbb{R}$ the substitution operator $\hat{f} : C \ni \phi \mapsto f \circ \phi \in C$ is continuous.

For a two-dimensional Banach space L the Jordan curve theorem asserts that the complement of the trace $|c|$ of a simple closed curve $c : [t_0, t_1] \to L$ consists of two connected components, a bounded one, int (c), which is called the interior of c, and an unbounded one, ext (c), called the exterior of c. We have

$$|c| = \partial \operatorname{int}(c) = \partial \operatorname{ext}(c).$$

If in addition c is continuously differentiable and if for some t the tangent vector $c'(t) = Dc(t)1$ and a vector $v \in L$ are linearly independent then there exists $\epsilon > 0$ so that the line segments $c(t) + (0, \epsilon)v$ and $c(t) + (-\epsilon, 0)v$ belong to different connected components of $L \setminus |c|$.

Solutions of a delay differential equation

$$\dot{x}(t) = g(t, x(t), x(t-1))$$

with $t_0 \in \mathbb{R}$ and $g : [t_0, \infty) \times \mathbb{R}^2 \to \mathbb{R}$ given, are continuous functions $x : [t_0 - 1, \infty) \to \mathbb{R}$ which are differentiable for $t > t_0$ and satisfy the differential equation for $t > t_0$. In case $g : \mathbb{R}^3 \to \mathbb{R}$ one also considers solutions which are defined and differentiable on the whole real line and satisfy the differential equation everywhere. For a given map $x : I \to \mathbb{R}$, $I \subset \mathbb{R}$, and $t \in \mathbb{R}$ with $[t-1, t] \subset I$ the segment $x_t : [-1, 0] \to \mathbb{R}$ is defined by $x_t(s) = x(t+s)$.

For a solution $x : \mathbb{R} \to \mathbb{R}$ of an autonomous delay differential equation

$$\dot{x}(t) = g(x(t), x(t-1))$$

with $g : \mathbb{R}^2 \to \mathbb{R}$ locally Lipschitz continuous the α-limit set $\alpha(x)$ is defined to consist of all accumulation points of all sequences $(x_{t_n})_1^\infty$ with $t_n \to -\infty$ as $n \to \infty$. In case the semi-orbit $\{x_t : t \leqslant 0\}$ has a compact closure the α-limit set is nonempty, compact, connected, and invariant in the sense that for every $\phi \in \alpha(x)$ there is a solution $y : \mathbb{R} \to \mathbb{R}$ with $y_0 = \phi$ and $y_t \in \alpha(x)$ for all $t \in \mathbb{R}$.

The Landau symbol $o(z)$ is used as an abbreviation for the values of a function on a neighborhood V of 0 in $\mathbb{R}^m$ or $\mathbb{R}^m \times \mathbb{R}^k$ which in the first case satisfies

$$o(0) = 0 \quad \text{and} \quad \frac{o(z)}{\|z\|} \to 0 \quad \text{as } 0 \neq z \to 0$$

and in the second case $o(0, \zeta) = 0$ for all $\zeta \in \mathbb{R}^k$ with $(0, \zeta) \in V$ and

$$\frac{o(z, \zeta)}{\|z\|} \to 0 \quad \text{as } z \to 0 \text{ uniformly with respect to } \zeta.$$

The general reference for results on delay differential equations which are used in the sequel is [5]. See also [7,9].

2. SOLUTIONS, COMPACTNESS, INJECTIVITY, FORWARD INVARIANCE

Proposition 2.1. *Let $t_1 > t_0$. Let $g : [t_0, t_1] \times \mathbb{R} \to \mathbb{R}$ be continuous and locally Lipschitz continuous with respect to the second variable. Suppose there exists $\eta_0 > 0$ with*

$$\eta g(t, \eta) < 0 \ \textit{for} \ |\eta| \geqslant \eta_0.$$

Then for every $y_0 \in \mathbb{R}$ the maximal solution of the initial value problem (IVP)

$$\dot{y} = g(t, y), \qquad y(t_0) = y_0$$

is defined on $[t_0, t_1]$.

Proof. Suppose the assertion is false. Then there exists $y_0 \in \mathbb{R}$ so that the associated maximal solution is defined on an interval $[t_0, t_e)$ with $t_0 < t_e \leqslant t_1$ and

$$\limsup_{t \nearrow t_e} |y(t)| = \infty.$$

On the other hand,

$$\frac{dy^2}{dt}(t) = 2y(t)\dot{y}(t) = 2y(t)g(t, y(t)) < 0$$

for all $t \in [t_0, t_e)$ with $|y(t)| \geqslant \eta_0$, which implies that y is bounded, in contradiction to the previous statement. □

For $a>0$ consider the nonlinearity $f:\mathbb{R}\ni\xi\mapsto(a-|\xi|)\xi\in\mathbb{R}$. Then Eq. (1.1) can be written as

$$\dot{x}(t)=f(x(t))-ax(t-1).$$

Corollary 2.1. *For every $\phi\in C$ there exists a uniquely determined solution $x^\phi:[-1,\infty)\to\mathbb{R}$ of Eq. (1.1) with $x_0^\phi=\phi$.*

Proof. Define $g_1:[0,1]\times\mathbb{R}\to\mathbb{R}$ by $g_1(t,\xi)=f(\xi)-a\phi(t-1)$. As $\lim_{\xi\to\infty}(\xi)=-\infty$ and $\lim_{\xi\to-\infty}f(\xi)=\infty$ we can apply Proposition 2.1 to $g=g_1$. This yields the restriction of x^ϕ to [0, 1]. Then proceed by induction, setting

$$g_{n+1}(t,\xi)=f(\xi)-ax^\phi(t-1)$$

for $n\leqslant t\leqslant n+1, n\in\mathbb{N}, \xi\in\mathbb{R}$. □

The equation

$$F(t,\phi)=x_t^\phi$$

define a continuous semiflow $F:[0,\infty)\times C\to C$. We have continuous dependence on initial conditions in the sense that given $\phi\in C, t_0\geqslant 0$ and $\epsilon>0$ there exists $\delta>0$ with

$$|x^\psi(t)-x^\phi(t)|<\epsilon \quad \text{on } [-1,\, t_0]$$

for all $\psi\in\phi+C_\delta$. Each map $F(t,\cdot):C\to C, t\geqslant 0$, is continuously differentiable, with

$$D_2F(t,\phi)\chi=v_t^{\phi,\chi}$$

given by the solution $v^{\phi,\chi}:[-1,\infty)\to\mathbb{R}$ of the variational equation along x^ϕ, i.e.,

$$\dot{v}(t)=f'(x^\phi(t))v(t)-av(t-1)$$

with initial condition $v_0=\chi$. Moreover, the restriction of F to $(1,\infty)\times C$ is continuously differentiable, with

$$D_1F(t,\phi)1=\dot{x}_t^\phi.$$

Corollary 2.2. *The map $F_1:C\ni\phi\mapsto F(1,\phi)\in C^1$ is continuous, and all maps $F(t,\cdot), t\geqslant 1$, are compact.*

Proof. By Eq. (1.1), $\dot{x}_1^\phi = \hat{f} \circ F(1,\phi) - a\phi$ for all $\phi \in C$. It follows that F_1 is continuous. Let $t \geqslant 1$. Then $F(t,\cdot) = F(t-1,\cdot) \circ F(1,\cdot) = F(t-1,\cdot) \circ J \circ F_1$ is compact. □

Proposition 2.2. *All maps $F(t,\cdot), t \geqslant 0$, and all derivatives $D_2F(t,\phi)$, $t \geqslant 0$ and $\phi \in C$, are injective.*

Proof. 1. Let $t \geqslant 0$ be given.

2. Consider the map $F(t,\cdot)$. Let $\phi \neq \psi$ in C be given. For all $s \in (-1,0)$ with $\phi(s) \neq \psi(s)$ and $x^\phi(s+1) = x^\psi(s+1)$ we have

$$\dot{x}^\phi(s+1) = f(x^\phi(s+1)) - a\phi(s) \neq f(x^\psi(s+1)) - a\psi(s) = \dot{x}^\psi(s+1).$$

This implies $x_1^\phi \neq x_1^\psi$. Uniqueness for the IVPs associated with Eq. (1.1) yields $x_s^\phi \neq x_s^\psi$ for all $s \in (0,1)$. Using induction one finds $x_s^\phi \neq x_s^\psi$ for all $s \geqslant 0$. In particular, $F(t,\phi) \neq F(t,\psi)$.

3. Let $\phi \in C$ be given. Consider the derivative $D_2F(t,\phi)$. Let $\chi \in C\backslash\{0\}$. It remains to show $D_2F(t,\phi)\chi \neq 0$. Recall $D_2F(t,\phi)\chi = v_t^{\phi,\chi}$. Set $v = v^{\phi,\chi}$. For all $s \in (-1,0)$ with $\chi(s) \neq 0$ and $v(s+1) = 0$ we have

$$\dot{v}(s+1) = f'(x^\phi(s+1))v(s+1) - a\chi(s) = -a\chi(s) \neq 0.$$

This implies $v_1 \neq 0$. Uniqueness for the IVPs associated with the variational equation along x^ϕ yields $v_s \neq 0$ for all $s \in (0,1)$. Using induction one finds $v_s \neq 0$ for all $s \geqslant 0$. In particular, $D_2F(t,\phi)\chi = v_t \neq 0$. □

Proposition 2.3. *For all $t \geqslant 0$, $F(t, C_{2a}) \subset C_{2a}$ and $F(t, \overline{C_{2a}}) \subset \overline{C_{2a}}$.*

Proof. Let $\phi \in C_{2a}$, $x = x^\phi$. Assume $|x(t)| \geqslant 2a$ for some $t > 0$. Then there exists $t_0 > 0$ with $|x(t_0)| = 2a > |x(t)|$ for $-1 \leqslant t < t_0$. In case $x(t_0) = 2a$,

$$0 \leqslant \dot{x}(t_0) = f(2a) - ax(t_0 - 1) < f(2a) + 2a^2 = 0,$$

which is a contradiction. The argument in case $x(t_0) = -2a$ is analogous. We infer $|x(t)| < 2a$ for all $t \geqslant 0$. Consequently, $F(t, C_{2a}) \subset C_{2a}$ for all $t \geqslant 0$. The remaining part of the assertion follows by continuity. □

We also have the following result, which will not be used in the sequel but seems of interest in itself.

Proposition 2.4. *For every $\phi \in C$,*

$$-2a \leqslant \liminf_{t\to\infty} x^\phi(t) \leqslant \limsup_{t\to\infty} x^\phi(t) \leqslant 2a.$$

Proof. 1. Assume there exists $\phi \in C$ wiith $F(t, \phi) \notin \overline{C_{2a}}$ for all $t \geqslant 0$. Set $x = x^{\phi}$. There is a sequence $(t_j)_0^{\infty}$ with $t_j \nearrow \infty$ as $j \to \infty$ and $|x(t_j)| > 2a$ for all integers $j \geqslant 0$.
Claim: For every $t > 0$ with $x(t) > 2a$ $(x(t) < -2a)$ there exists $s > t$ with $x(t) = 2a$ $(x(t) = -2a)$.
Proof. Let $x(t) > 2a$. If $x(s) > 2a$ on $[t, \infty)$ then

$$\dot{x}(s) \leqslant f(2a) - a2a = -4a^2 < 0$$

on $[t+1, \infty)$, which yields a contradiction.

2. We may assume $x(t_j) > 2a$ for all integers $j \geqslant 0$, or $x(t_j) < -2a$ for all integers $j \geqslant 0$. Consider the first case.
3. It follows that there is a sequence of local maxima m_j, $j \in \mathbb{N}_0$, of x with $m_j \to \infty$ and $2a < x(m_j) \to \limsup_{t\to\infty} x(t) \leqslant \infty$. From Eq. (1.1),

$$0 = \dot{x}(m_j) = f(x(m_j)) - ax(m_j - 1).$$

That is for $\xi = x(m_j)$ and $\eta = x(m_j - 1)$ we have $\xi > 2a$ and $0 > a\eta = f(\xi) = a\xi - \xi^2$. Notice that $a\xi - \xi^2 < -a\xi$ since $2a < \xi$. Hence

$$x(m_j - 1) = \eta < -\xi = -x(m_j).$$

For $s_j = m_j - 1$, we have $x(s_j) < -2a$, and $s_j \to \infty$ as $j \to \infty$. There is a sequence $(s_{j*})_0^{\infty}$ with $s_{j*} > s_j$ and $x(s_{j*}) = -2a$ for all integers $j \geqslant 0$. It follows that there is a sequence $(M_j)_0^{\infty}$ of local minima of x with $x(M_j) < -2a$ for all integers $j \geqslant 0$, and $M_j \to \infty$ and $x(M_j) \to \liminf_{t\to\infty} x(t) \geqslant -\infty$ as $j \to \infty$.
4. Proof that x is bounded. Otherwise there exists a local extremum $m > 0$ of x so that $|x(t)| < |x(m)|$ for $-1 \leqslant t < m$ and $2a < |x(m)|$. Arguing as in part 3 we find $|x(m-1)| > |x(m)|$, a contradiction.
5. It follows that $c = \limsup_{t\to\infty} x(t) \geqslant 2a$ and $d = \liminf_{t\to\infty} x(t) \leqslant -2a$ are finite. As $x(m_j - 1) < -x(m_j)$ (see part 3) we infer $d \leqslant -c$. Using the sequence $(M_j)_0^{\infty}$ of local minima of x and analogous arguments we obtain $-c \leqslant d$. Together, $-c = d$. Hence

$$\begin{aligned} -c &= \liminf_{t\to\infty} x(t) \leqslant \liminf_{j\to\infty} x(M_j - 1) \\ &= \liminf_{j\to\infty} \left(x(M_j) - \frac{x(M_j)^2}{a}\right) \text{(see part3)} \\ &= c - \frac{c^2}{a}, \quad \text{or } c \leqslant 2a. \end{aligned}$$

6. The proof in the other case in part 2 is analogous. □

3. SLOWLY OSCILLATING SOLUTIONS

We call a function $x : I \to \mathbb{R}, I \subset \mathbb{R}$, slowly oscillating if for any pair of zeros $z_* > z$ we have $z_* > z + 1$. Soon we shall see that slowly oscillating solutions of Eq. (1.1) are abundant. It is convenient to begin with solutions of the more general, nonautonomous equation

$$\dot{y}(t) = h(t, y(t)) - ay(t-1) \tag{3.1}$$

with $a > 0$ and a continous function $h : [0, \infty) \times \mathbb{R} \to \mathbb{R}$ which is locally Lipschitz continous with respect to the second variable and satisfies

$$h(t, 0) = 0 \quad \text{for all } t \geqslant 0$$

and

$$h(t, \eta) \to \infty (\to -\infty) \quad \text{as } \eta \to -\infty (\to \infty)$$

uniformly with respect to t in compact sets.

Example. *For a given solution $x : [-1, \infty) \to \mathbb{R}$ of Eq. (1.1) set $h(t, \eta) = f(\eta + x(t)) - f(x(t))$.*

Notice that for h as above, for $t_0 \geqslant 0$, and for every continuous function $c : [t_0, t_0 + 1] \to \mathbb{R}$ the map $g : [t_0, t_0 + 1] \times \mathbb{R} \to \mathbb{R}$ defined by $g(t, \eta) = h(t, \eta) - c(t)$ satisfies the hypotheses of Proposition 2.1.

Segment of slowly oscillating functions belong to the set $Z \subset C$ of data with at most one zero. The closure $S_1 = \bar{Z}$ is the set of all $\phi \in C$ which are nonnegative, or nonpositive, or have a zero $z \in (-1, 0)$ so that for some $j \in \{0, 1\}$,

$$(-1)^j \phi(t) \geqslant 0 \text{ on } [-1, z], (-1)^j \phi(t) \leqslant 0 \quad \text{on } [z, 0].$$

That is, S_1 is the set of data with at most one sign change. One can show that $S_1 \setminus \{0\}$ is homotopy equivalent to the unit circle S^1, but we shall not make use of this.

Proposition 3.1. *Let $y : [-1, \infty) \to \mathbb{R}$ be a solution of Eq. (3.1).*

(i) If $0 \neq y_0 \in S_1$ then $0 \neq y_t \in S_1$ for $0 \leqslant t \leqslant 3, 0 \notin y([t-1, t])$ for some $t \in [0, 3]$, and $y_t \in Z$ for $t \geqslant 3$.

(ii) If $y(z)=0$ for some $z \geqslant 0$ and $0 \notin y([z-1, z))$ then

$$\operatorname{sign}(y(t)) = -\operatorname{sign}(y(z-1)) \neq 0 \quad \text{for } z < t \leqslant z+1.$$

Proof. 1. Proof of $0 \neq y_t \in S_1$ for $0 \leqslant t \leqslant 1$ and $0 \notin y([t-1, t])$ for some $t \in [0, 2]$ in case (A) that there exists $z \in (-1, 0]$ with $0 \leqslant y(t)$ in $[-1, z]$, $0 < y(t_+)$ for some $t_+ \in (-1, z)$, and $y(t) \leqslant 0$ in [z, 0].

1.1 Proof of $y(t) \leqslant 0$ in [0, z+1]. We compare y to the solution $w : [0, z+1] \to \mathbb{R}$ of the IVP

$$\dot{w} = h(t, w) - ay(t-1), \quad w(0) = 0.$$

As solutions can not cross, we get $y(t) \leqslant w(t)$ on [0, z+1], and it remains to show $w(t) \leqslant 0$ on [0, z+1]. We have

$$w(t) = \lim_{\epsilon \searrow 0} w_\epsilon(t) \text{ on } [0, z+1]$$

with the solutions $w_\varepsilon : [0, z+1] \to \mathbb{R}$ to the IVPs

$$\dot{w} = h(t, w) - ay(t-1) - \epsilon, \quad w(0) = 0.$$

For all $\epsilon > 0$ and $t \in [0, z+1]$,

$$w_\epsilon(t) \leqslant 0$$

since otherwise we obtain from $w_\epsilon(0) = 0 > -\epsilon \geqslant 0 - ay(-1) - \epsilon = \dot{w}_\epsilon(0)$ some $t \in (0, z+1)$ with $w_\epsilon(t) = 0$ and $\dot{w}_\epsilon(t) \geqslant 0$, in contradiction to

$$\dot{w}_\epsilon(t) = 0 - ay(t-1) - \epsilon \leqslant -\epsilon < 0.$$

1.2 Proof of $y(t) < 0$ for some $t \in (0, z+1]$. Otherwise, $y(t) = 0$ on [0, z+1], and consequently $\dot{y}(t_+ + 1) = 0 - ay(t_+) < 0$, which yields a contradiction.

1.3 Claim: If $y(t_1) < 0$ and $0 \leqslant t_1 < z+1$ then $y(t) < 0$ on $[t_1, z+1]$. Proof. On $[t_1, z+1]$, $y(t) < w(t)$ for the solution $w : [t_1, z+1] \to \mathbb{R}$ of the IVP

$$\dot{w} = h(t, w) - ay(t-1), \quad w(0) = 0.$$

As in part 1.1 one finds $w(t) \leqslant 0$ in $[t_1, z+1]$.

1.4 It follows that either (A1) $y(t) < 0$ on $[0, 1]$, or (A2) there exists $z_* \in [0, z+1)$ with $y(t) = 0$ in $[0, z_*]$ and $y(t) < 0$ in $(z_*, z+1]$. In case (A1) the assertion $0 \neq Y_t \in S_1$ for $0 \leqslant t \leqslant 1$ is obvious. Also, for $\epsilon > 0$ sufficiently small the restriction of y to $[\epsilon, 1+\epsilon]$ has no

zero. In case (A2), we find for $t \in [z+1, z_*+1]$ that $y(t) < w(t)$, with the solution $w : [z+1, z_*+1] \to \mathbb{R}$ of the IVP

$$\dot{w} = h(t, w) - ay(t-1) = h(t, w), \quad w(z+1) = 0,$$

i.e., $w(t) = 0$ on $[z+1, z_*+1]$. Hence $y(t) < 0$ in $(z_*, z_*+1]$. As in case (A1) we see that for $t \in [0, 1], 0 \neq y_t \in S_1$, and for $\epsilon > 0$ sufficiently small, $0 \notin y([z_*+\epsilon, z_*+\epsilon+1])$.

2. In case that $-y_0$ has the properties stated in (A) arguments as above yield $0 \neq y_t \in S_1$ for $0 \leqslant t \leqslant 1$ and $0 \notin y([t-1, t])$ for some $t \in [0, 2]$.
3. In case $0 \leqslant y(t)$ in $[-1, 0]$ and $0 < y(0)$, either $0 < y(t)$ on $[0, 1]$, or there is a smallest zero $z \in (0, 1]$ of y. Then y_z has property (A). Now it becomes obvious how to complete the proof that $0 \neq y_0 \in S_1$ implies $0 \neq y_t \in S_1$ for $0 \leqslant t \leqslant 1$ and $0 \notin y([t_*-1, t_*])$ for some $t_* \in [0, 3]$. Using induction on derives $0 \neq y_t \in S_1$ for $0 \leqslant t \leqslant 3$.
4. Proof of (ii). Suppose $0 \leqslant z, 0 < y(t)$ in $[z-1, z)$, and $y(z) = 0$. Then $\dot{y}(z) < 0$. Hence $y(t) < 0$ on $(z, z+\epsilon]$ for some $\epsilon \in (0, 1]$, and $y(t) < w(t)$ in $[z+(\epsilon/2), z+1]$ for the solution $w : [z+(\epsilon/2), z+1] \to \mathbb{R}$ of the IVP

$$\dot{w} = h(t, w) - ay(t-1), \quad w(z+\frac{\epsilon}{2}) = 0.$$

As in part 1.1 we have $w(t) \leqslant 0$ on $[z+(\epsilon/2), z+1]$. It follows that $y(t) < 0$ in $(z, z+1]$. The proof in case $y(t) < 0 = y(z)$ in $[z-1, z)$ is analogous.
5. Using (ii) and the result of part 3 one shows that for any zero $z > t_*$ of y, $0 \notin y([z-1, z) \cup (z, z+1])$. This yields $y_t \in Z$ for $t \geqslant 3$. □

In other words, nonzero initial data in $S_1 = \bar{Z}$ yield solutions (of Eq. (1.1)) which become slowly oscillating after finite time. Notice that we did not show that solutions actually have zeros in $(0, \infty)$. In particular, we do not yet know whether there are slowly oscillating solutions of Eq. (1.1) with infinitely many zeros. The proof that this is indeed the case for $a > 1$ will be given in Section 6. It is not as elementary as the arguments in Section 3. We shall employ the following preliminary result.

Proposition 3.2. *Suppose $\phi \in \overline{C_{2a}}$ and the solution $x = x^\phi$ of Eq. (1.1) has no zero. Then*

$$x(t) \to 0 \quad as \quad t \to \infty.$$

Proof. 1. Let $\phi(0) > 0$. Proposition 2.3 shows that $c = \limsup_{t\to\infty} x(t)$ and $d = \liminf_{t\to\infty} x(t)$ satisfy

$$0 \leqslant d \leqslant c \leqslant 2a.$$

2. Proof of $d = c$. Assume $d < c$.

2.1 In case $0 < d$ choose $\epsilon > 0$ with

$$2a\epsilon - (d-\epsilon)^2 < 0.$$

There exists $t \geqslant 0$ with $d - \epsilon < x(s)$ for all $s \geqslant t$. The inequality $d < c$ implies that there is a local minimum $t_m > t + 1$ of x with $x(t_m) < d + \epsilon$. Hence

$$\begin{aligned} 0 = \dot{x}(t_m) &= a(x(t_m) - x(t_m - 1)) - (x(t_m))^2 \\ &\leqslant a(d + \epsilon - (d - \epsilon)) - (d - \epsilon)^2 = 2a\epsilon - (d-\epsilon)^2 < 0, \end{aligned}$$

which is a contradiction.

2.2 In case $d = 0$ choose $m \in (0, c)$ with $m < \phi(t)$ for $-1 \leqslant t \leqslant 0$. As $d = 0$ we find $t > 0$ with $x(t) = m$ and $x(s) > m$ for $0 \leqslant s < t$. Then $\dot{x}(t) \leqslant 0$. Using $m < c$ we find a smallest zero $t_m \geqslant t$ of $\dot{x}$. Either $t_m = t$, or $t < t_m$ and $\dot{x}(s) < 0$ on $[t, t_m)$. In both cases,

$$x(t_m - 1) > x(t_m) \geqslant 0.$$

Hence

$$0 = \dot{x}(t_m) = a(x(t_m) - x(t_m - 1)) - (x(t_m))^2 < -(x(t_m))^2 \leqslant 0,$$

which is a contradiction.

3. We have shown $x(t) \to c \geqslant 0$ as $t \to \infty$. Consequently,

$$\lim_{t\to\infty} \dot{x}(t) = a(c - c) - c^2 = -c^2.$$

In case $c > 0$ we obtain a contradiction.

4. The proof in cas $\phi(0) < 0$ is analogous. □

Remark 3.1. With regard to the hypothesis in Proposition 3.2, notice that in case $\phi \in \overline{C_{2a}}$ has no zero and $x = x^\phi$ does not change sign there is no zero of x at all since a first zero would be simple, due to Eq. (1.1).

4. LINEARIZATION

The solution operators $T_t = D_2 F(t, 0), t \geqslant 0$, of the variational equation

$$\dot{v}(t) = a(v(t) - v(t-1)) \tag{4.1}$$

along the zero solution of Eq. (1.1) form a strongly continuous semi-group whose generator $G : D \to C$ is given by

$$D = \{\phi \in C^1 : \dot{\phi}(0) = a(\phi(0) - \phi(-1))\}$$

and

$$G\phi = \dot{\phi}$$

The spectrum of G consists of eigenvalues which are isolated points. They are given by the characteristic equation

$$\lambda - a(1 - e^{-\lambda}) = 0. \tag{4.2}$$

The order of a solution λ of Eq. (2) coincides with the dimension of the generalized eigenspace of λ considered as a point in the spectrum of the complexification of G.

Proposition 4.1. *For every $a > 0$, $\lambda = 0$ is a soltion of Eq. (4.2), simple for $a \neq 1$ and double for $a = 1$. For $a \neq 1$ there exists a unique real solution $u \neq 0$ of Eq. (4.2); u is simple, and $u < 0$ for $0 < a < 1$, $0 < u$ for $1 < a$. For $a < 1$ all other solutions of Eq. (4.2) satisfy $Re(\lambda) < u$. For $1 \leqslant a$ all other solutions of Eq. (4.2) satisfy $Re(\lambda) < 0$.*

Proof. A number $\lambda \in \mathbb{C}$ is a solution of Eq. (4.2) if and only if $\zeta = \lambda - a$ is a solution of the equation

$$\zeta + ae^{-\zeta} = 0 \tag{4.3}$$

with $\alpha = ae^{-a}$. Notice that the function

$$g : (0, \infty) \ni a \mapsto ae^{-a} \in \mathbb{R}$$

has a strict global maximum at $a = 1$, with $g(1) = 1/e$. Use the results about Eq. (4.3) for $0 < \alpha \leqslant 1/e$ in, e.g. [20]. □

For every $a > 0$ let $C_0 \subset C$ denote the center space, i.e., realified generalized eigenspace of G given by the zero eigenvalue, and let L denote the realified generalized eigenspace of G given by the two leading real eigenvalues. For $a > 1$ let C_+ denote the unstable space, i.e., the realified generalized eigenspace of G given by the positive eigenvalue.

Corollary 4.1. For all $a>0$, dim $L=2$ and $C_0 \subset L$. For $0<a\neq 1$, dim $C_0=1$, while for $a=1$, dim $C_0=2$ and $C_0=L$. For $a>1$, dim$C_+=1$ and $L=C_0\oplus C_+$.

For $0<a\neq 1$, $C_0=\mathbb{R}\eta_0$ where $\eta_0(t)\equiv 1$. For $a=1$, $L=C_0=\mathbb{R}\eta_0\oplus\mathbb{R}\sigma$, with $\sigma(t)=t$. For $a>1$, $C_+=\mathbb{R}\eta_u$ where $\eta_u(t)=e^{ut}$. For $0<a\neq 1$, $L=\mathbb{R}\eta_0\oplus\mathbb{R}\eta_u$.

Proof. Verify $G\eta_0=0$, $G\sigma=\eta_0$ in case $a=1$ and $G\eta_0=0$, $G\eta_u=u\eta_u$ for $0<a\neq 1$. □

Corollary 4.2. *For $a>1$ the zero solution of the lineraized Eq. (1) is unstable whereas for $0<a\leqslant 1$ it is stable but not asymptotically.*

For $0<a\neq 1$ the space L consists of the segments of the solutions

$$\mathbb{R}\ni t\mapsto \alpha+\beta e^{ut}\in\mathbb{R},\ (\alpha,\beta)\in\mathbb{R}^2$$

of Eq. (4.1). These solutions are constant or strictly monotone, with at most one zero in case $(\alpha,\beta)\neq(0,0)$. In particular,

$$L\subset Z\cup\{0\}.$$

The last inclusion holds for $a=1$ as well.

Both for the local and the global behavior of solutions the projections of the dynamics of the equation to the subspaces L and C_0 is crucial. To this end we employ two kinds of projections.

To study the local behavior we work with the spectral projection $\Pi: C\to C$ onto C_0. According to [7] it can be computed using the bilinear form on the product

$$C([0,1],\mathbb{C})\times C([-1,0],\mathbb{C}),$$

which is given by

$$\langle\psi,\phi\rangle=\langle\psi(0),\qquad \phi(0)\rangle-a\int_{-1}^{0}\psi(\xi+1)\phi(\xi)\,d\xi. \tag{4.4}$$

Let $C_0^*\subset C([0,1],\mathbb{R})$ denote the realified generalized eigenspace of the zero eigenvalue of the characteristic equation of the formally adjoint equation

$$\dot w(t)=a(w(t+1)-w(t)) \tag{4.5}$$

(see [7]). We have dim $C_0^*=1$ for $0<a\neq 1$ and dim $C_0^*=2$ for $a=1$. The projection Π is given by

$$\Pi(\chi)=<\Psi,\quad \chi>\Phi, \tag{4.6}$$

where Φ is the basis (η_0) of C_0 in case $0 < a \neq 1$ and $\Phi = (\eta_0 \sigma)$ for $a = 1$, and Ψ is a basis of the space C_0^* which is normalized so that $< \Psi, \Phi >$ is the unit matrix. That is, in case $0 < a \neq 1$, $\Psi = (\psi_0)$ and

$$\langle \psi_0, \eta_0 \rangle = 1 \tag{4.7}$$

while for $a = 1$, $\Psi = (\psi_0, \psi_1)$ and

$$\langle \psi_0, \eta_0 \rangle = 1, \quad \langle \psi_1, \sigma \rangle = 1, \quad \langle \psi_0, \sigma \rangle = \langle \psi_1, \eta_0 \rangle = 0. \tag{4.8}$$

The normalized basis Ψ can be obtained from any basis $\tilde{\Psi}$ of C_0^* by

$$\Psi = \langle \tilde{\Psi}, \Phi \rangle^{-1} \tilde{\Psi}, \tag{4.9}$$

see [7].

To study the global dynamics we choose a closed complementary space N for L in C in such a way that the associated projection P onto L makes it easy to describe the spiraling motion of projected flowlines $t \mapsto P_{x_t} \in L$, for slowly oscillating solutions x to Eq. (1.1). Define N to be the intersection of the closed hyperplanes

$$H_- = \{\phi \in C : \phi(-1) = 0\} \quad \text{and} \quad H_0 = \{\phi \in C : \phi(0) = 0\}.$$

We have codim $N = 2$ and

$$Z \cap N = \emptyset.$$

Proposition 4.2. $C = L \oplus N$.

Proof. As each $\chi \in L \setminus \{0\}$ has at most one zero we have $L \cap N = \{0\}$. Then the result follows from dim $L = 2 = \operatorname{codim} N$. □

There exists $\phi_0 \in (H_0 \cap L) \setminus H_-$ with $\phi_0(-1) = 1$ and $\phi_- \in (H_- \cap L) \setminus H_0$ with $\phi_-(0) = 1$. ϕ_0 and ϕ_- are linearly independent. Let $P : C \to C$ denote the projection along N onto L.

Proposition 4.3. *For every* $\phi \in C$, $P\phi = \phi(-1)\phi_0 + \phi(0)\phi_-$.

Proof. Let $\phi \in C$. Then $\phi = P\phi + \psi$ with $\psi \in N$. Hence $\phi(0) = (P\phi)(0) + \psi(0) = (P\phi)(0)$ and $\phi(-1) = (P\phi)(-1) + \psi(-1) = (P\phi)(-1)$. From $P\phi = c_0\phi_0 + c_-\phi_-$ with reals c_0, c_-,

$$(P\phi)(0) = c_0 \cdot 0 + c_- \cdot 1 = c_-,$$
$$(P\phi)(-1) = c_0 \cdot 1 + c_- \cdot 0 = c_0.$$

□

Slowly oscilatting solution behavior is now reflected in the projected flowlines as follows: For differntial functions $x : \mathbb{R} \to \mathbb{R}$ with a simple zero z, no zero in$[z-1, z) \cup (z, z+1]$, and a consecutive zero $\zeta > z+1$ we have in case $\dot{x}(z) > 0$ that the projected segment$Px_z$is on the half-axis$(-\infty, 0)\phi_0$. For $z < t < z+1, Px_t$ belongs to the open sector$(-\infty, 0)\phi_0 + (0, \infty)\phi_-$. Next,$Px_{z+1} \in (0, \infty)\phi_-$,and for $z+1 < t < \zeta$, $Px_t \in (0, \infty)\phi_0 + (0, \infty)\phi_-$.

Proposition 4.4. $PH_0 = \mathbb{R}\phi_0$, $P^{-1}(PH_0) = H_0$, $PH_- = \mathbb{R}\phi_-$, $P^{-1}(PH_-) = H_-$.

Proof. From $\phi_0 \in (H_0 \cap L) \setminus H_- \subset H_0 \setminus N$ we infer $H_0 = \mathbb{R}\phi_0 \oplus N$; it follows that $PH_0 = \mathbb{R}\phi_0$.

Suppose $P^{-1}(PH_0) \not\subset H_0$. Then $P\phi \in PH_0$ for some $\phi \in C \setminus H_0$.Consequently ,$P^{-1}(PH_0) = C$, which implies $L = PC \subset PH_0$, in contradiction to dim $L = 2$ and $PH_0 = \mathbb{R}\phi_0$. - It follow that $P^{-1}(PH_0) = H_0$.

The assertion for H_- proved analogously. □

5. CENTER MANIFOLDS AND STABILITY OF EQUILIBRIUM

As general reference for invariant manifold theory in case of delay differential equations (see[5, 7, 8]).

By proposition 4.1, zero is a root of the characteristic equation of Eq. (4.1) for all $a > 0$. Therefore, stability of the eauilibrium 0 of the nonlinear Eq. (1.1) is not decided by its linearization (4.1) for any $a \in (0, 1]$. In order to resolve the stability question we employ that standard tool for the study of stability in such a case – the center manifold reduction.

Recall that the center manifold W_c of the equilibrium $0 \in C$ is a C^1- submanifold of C of dimension $m =$ dim C_0 which contains 0, is tangent to the center space C_0 at 0, and is locally positively invariant under the semiflow. The latter means that there is a neighbourhood U of 0 in C so that for every $\phi \in W_c$ and $t \geqslant 0$ with $F([0, t] \times \{\phi\}) \subset U$ we have $F(t, 0) \in W_c$.

Remark 5.1. In fact, the center manifold is not uniquely defined in general. Still, we can speak about a single, arbitrarily picked manifold since the semiflows on all the center manifolds are conjugate. That is, any two center manifolds admit a homeomorphism mapping trajectories onto trajectories.

The nonlinear term in Eq. (1.1) being C^1-smooth, by [5] Section VII.6, Theorem IX.5.3 and Corollary IX.7.8, a center manifold exists and is tangent to C_0 at 0.

By the center manifold reduction we understand the restriction of the local semiflow of Eq. (1.1) to the center manifold. This restriction is a local flow generated by a system of $m = dim\, W_c$ first order ordinary

differential equations. In our case, $m=1$ for $0<a\neq 1$ and $m=2$ for $a=1$. We will freely use the term reduction for this system as well.

The center manifold is given by the formula

$$W_c=\{\phi\in C:\phi=\Phi z+h(z), z \text{ in a neighbourhood of zero in } \mathbb{R}^m\}$$

with Φ from Section 4 each h a C^1-map from $\mathbb{R}^m$ to the complementary realified generalized eigenspace of C_0 such that $h(z)=0(z)$.

The center manifold reduction for the case $a=1$ is computed in [5], Section IX.10. Although the construction can be extended to the even simpler case $0<a\neq 1$, for the convenience of the reader we have chosen to present an outline of an alternative construction for both cases. Employing the formal adjoint, this computation follows [7, 8] and appears to require less effort.

To this end we write Eq. (1.1) in the form

$$\dot{x}=\Lambda x_t+q(x_t),$$

where $\Lambda: C\to\mathbb{R}$ and $q: C\to\mathbb{R}$ are given by

$$\Lambda\phi=a\phi(0)-a\phi(-1), \quad \text{and} \quad q(\phi)=-|\phi(0)|\phi(0). \tag{5.1}$$

The reduction of the semiflow to the center manifold is the equation

$$\dot{z}=Bz+cq(\Phi z+h(z)), \tag{5.2}$$

B given by

$$\Lambda\Phi=\Phi B \tag{5.3}$$

and $c=\Psi(0)$,Ψ introduced in Section 4.

We first deal with the case $0<a\neq 1$. Although for $a>1$ the center manifold reduction is not needed for stability (by Corollary 4.2, 0 is unstable by linearization in this case) it turns out to be important for the proof of the existence of a periodic solution in Section 7.

For $0<a\neq 1$, the dimension of the center manifold is 1. The subspace $C_0^*\subset([0,1],\mathbb{R})$ is spanned by the function $\tilde{\psi}_0$ given by

$$\tilde{\psi}_0(\theta)=1 \quad \text{for } 0\leqslant\theta\leqslant 1.$$

The formula for the center manifold reads

$$W_c=\{\phi\in C:\phi=z\eta_0+h(z), z \text{ in a neighborhood of zero in } \mathbb{R}\}.$$

By (4.9) the normalized vector ψ_0 in C_0^* satisfies

$$\psi_0(\theta)=\Big[\langle\tilde{\psi}_0,\eta_0\rangle\Big]^{-1}\tilde{\psi}_0(\theta)=\Big[\langle\tilde{\psi},\eta_o\rangle\Big]^{-1}, \tag{5.4}$$

i.e.,

$$\psi_0(\theta) = \left[\tilde{\psi}_0(0)\eta_0(0) - a\int_{-1}^{0}\tilde{\psi}_0(\xi+1)\eta_0(\xi)d\xi\right]^{-1}$$
$$= \left[1 - a\int_{-1}^{0}d\xi\right]^{-1} = \frac{1}{1-a}. \tag{5.5}$$

for $0 \leqslant \theta \leqslant 1$. Hence, in this case the center manifold reduction (5.2) reads

$$\dot{z} = -\frac{1}{1-a}q(z\eta_0 + h(z)) = -\frac{1}{1-a}|z|z + o(z^2). \tag{5.6}$$

Theorem 5.1. *The equilibrium solution of Eq. (5.6) is locally asymptotically stable for $a < 1$ and unstable for $a > 1$.*

Proof. The conclusion follows immediately from the sign of the leading term of Eq. (5.6). □

Remark 5.2. The last statement in the theorem is somewhat surprising. By it, the seemingly stabilizing nonlinear term projects for $a > 1$ to a repelling force in the center manifold. This turns out to be important for the global dynamics studied in the following chapters.

By [5], Section IX.8, we have the folowing.

Corollary 5.1. *For $0 < a < 1$ the stationary point $0 \in C$ of Eq. (1.1) is asymptotically stable.*

We now turn to a discussion of the critical case $a = 1$. In this case, the dimension of the center manifold is 2. By Corollary 4.1, the eigenspaces C_0 is spanned by the functions η_0 and σ, and the points of C_0 can be represented by $z_1\eta_0 + z_2\sigma$ with $z_1, z_2 \in \mathbb{R}$ Since $\dot{\sigma} = \eta_0$ and $\dot{\eta}_0 = 0$, the restriction of the linear Eq. (4.1) to C_0 given by the system $\dot{z} = B_z$ with

$$B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{5.7}$$

Let $\tilde{\Psi} = (\tilde{\psi}_1, \tilde{\psi}_2)$ be the basis of C_0^* given by

$$\tilde{\psi}_1(\theta) = 1, \qquad \tilde{\psi}_2(\theta) = \theta \quad \text{for } 0 \leqslant \theta \leqslant 1.$$

By definition (4.4) of the scalar product we have

$$\langle \tilde{\psi}_1, \eta_0 \rangle = 1 - \int_{-1}^{0} d\xi = 0,$$

$$\langle \tilde{\psi}_2, \eta_0 \rangle = 0 - \int_{-1}^{0} (\xi + 1)\, d\xi = -\frac{1}{2},$$

$$\langle \tilde{\psi}_1, \sigma \rangle = 0 - \int_{-1}^{0} \xi\, d\xi = \frac{1}{2},$$

$$\langle \tilde{\psi}_2, \sigma \rangle = 0 - \int_{-1}^{0} (\xi + 1)\xi\, d\xi = -\frac{1}{6}. \tag{5.8}$$

Substituting (5.8) into (4.9) we obtain

$$\Psi = \left[\langle \tilde{\Psi}, (\eta_u, \sigma) \rangle\right]^{-1} \tilde{\Psi} = \begin{pmatrix} 0 & 1/2 \\ -1/2 & 1/6 \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\psi}_1 \\ \tilde{\psi}_2 \end{pmatrix}.$$

The normalized basis Ψ is therefore given by

$$\Psi(\theta) = \begin{pmatrix} 2/3 & -2 \\ 2 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \theta \end{pmatrix} = \begin{pmatrix} 2/3 - 2\theta \\ 2 \end{pmatrix}. \tag{5.9}$$

As a conclusion we obtain

$$c = \Psi(0) = \begin{pmatrix} 2/3 \\ 2 \end{pmatrix}. \tag{5.10}$$

Let $z = (z_1, z_2)$. Then we have

$$((\eta_0, \sigma)z)(\theta) = z_1 \eta_0(\theta) + z_2 \sigma(\theta) = z_1 + \theta z_2 \quad \text{for } -1 \leqslant \theta \leqslant 0. \tag{5.11}$$

Hence, the reduction to the center manifold is

$$\begin{aligned} \dot{z}_1 &= z_2 - \frac{2}{3}|z_1| z_1 + o((|z_1| + |z_2|)^2), \\ \dot{z}_2 &= -2|z_1| z_1 + o((|z_1| + |z_2|)^2). \end{aligned} \tag{5.12}$$

As we have mentioned in Section 1, numerical simulations indicate that the equilibrium solution $x \equiv 0$ is globally asymptotically stable, the nonzero solutions oscillating around zero with the distance of zeros increasing. Below, we present an incomplete collection of results about the local dynamics supporting partially these observations. Unlike in the case

$0 < a \neq 1$, for $a = 1$ we are currently not able to give a complete rigorous local analysis.

Let $z = (z_1, z_2)$ be a solution of the system (5.12). Substituting polar cocordinates

$$z_1 = r \sin \omega, \qquad z_2 = r \cos \omega$$

We find the system of equations

$$\dot{r} = r \cos \omega \sin \omega - \frac{2}{3} r^2 |\sin \omega| \sin \omega (\sin \omega + 3 \cos \omega) + o(r)^2, \tag{5.13}$$

$$\dot{\omega} = \cos^2 \omega + \frac{2}{3} r |\sin \omega| \sin \omega (3 \sin \omega - \cos \omega) + o(r) \tag{5.14}$$

for the real functions r and ω.

Proposition 5.1. *There exists $r_0 > 0$ so that for any solution (r, ω) : $[o, \infty) \to [0, \infty) \times \mathbb{R}$ of the system (5.13) and (5.14) with sup $r < r_0$ we have $\omega(t) \to \infty$ as $t \to \infty$. If in addition $\lim_{t\to\infty} r(t) = 0$ then $\lim_{t\to\infty} \tau(t) = \infty$ for any function $\tau : [0, \infty) \to \mathbb{R}^+$ satisfying*

$$\omega(t + \tau(t)) = \omega(t) + 2\pi \quad \textit{for all } t \geqslant 0.$$

Proof. Denote $M = \{\gamma : |\gamma - k\pi| \leqslant \pi/4$ for some integer $k\}$, $N =$ cl $(\mathbb{R} \backslash M)$. We have

$$|\sin \omega| \leqslant 1/\sqrt{2} \leqslant |\cos \omega| \quad \text{for } t \in M, \tag{5.15}$$

$$|\cos \omega| \leqslant 1/\sqrt{2} \leqslant |\sin \omega| \quad \text{for } t \in N. \tag{5.16}$$

obviously, $\omega^{-1}(M)$is a union of disjoint closed intervals separated by open intervals of $\omega^{-1}(\text{int } N)$. More precisely, if $[\tau_1, \tau_2]$ and $[\tau_3, \tau_4]$,$\tau_2 < \tau_3$ are two consecutive disjoint intervals of $\omega^{-1}(M)$ then $[\tau_2, \tau_3]$ is an interval of $\omega^{-1}(N)$. A similar statement holds with M and N interchanged. If $r(t)$is bounded, the lengths of the intervals of $\omega^{-1}(M)$ are bounded from below. To prove the first part of the proposition we show that if $r(t)$ stays sufficiently small and $\omega(t_1) \in M(N)$, then there exists a $t_2 > t_1$ for which $\omega(t_2) \notin M(N,$ respectively). By (5.15), for $\omega(t) \in M$ we have

$$\dot{\omega}(t) \geqslant \cos^2 \omega(t) + o(r(t)) \geqslant \frac{1}{2} + o(r(t)). \tag{5.17}$$

Hence, for $r(t)$ sufficiently small and $\omega(t_1) \in M$ the assumption $\omega(t) \in M$ for $t \geqslant t_1$ leads to a contradiction.

For $\omega(t) \in N$ we have

$$\dot{\omega}(t) \geqslant \frac{4}{3} r(t) 2^{-3/2} + o(r(t)) > \beta r(t) \tag{5.18}$$

for some $\beta > 0$ provided r_1 was chosen sufficiently small. Thus, $\omega(t)$ is increasing in N. Let $\omega(t_1) \in N$. should $\omega(t)$ not leave N for some $t_g e t_1$, there would exist an $\omega^* \in N$ such that $\omega(t) \nearrow \omega^*$ for $t \to \infty$. We complete the first part of the proof by showing this to be impossible. Without loss of generality we assume $\omega^* \in (\pi/4, 3\pi/4]$. The leading two terms on the right-hand sides of (5.13) and (5.14) being periodic with period π, the proof applies to the cases $k \neq 0$ as well.

We first prove that $\omega^* \leqslant \pi/2$ leads to a contradiction. Indeed, since $\dot{\omega}(t) > 0$ in N, we can reparametrize t by r. That is, there is a smooth function $\tau : [\omega(t_1), \omega^*) \to \mathbb{R}$ such that $\tilde{r} = r \circ \tau$ satisfies

$$\frac{d\tilde{r}}{d\omega}(\omega) = \frac{(dr/dt)(\tau(\omega))}{(d\omega/dt)(\tau(\omega))} = \tilde{r}(\omega) \tan \omega + o(\tilde{r}(\omega)) > 0 \tag{5.19}$$

for $\tilde{r}$ sufficiently small. Hence, $\tilde{r}(\omega) \geqslant \tilde{r}(\omega(t_1))$ for all $\omega < \omega^*$, and, consequently, also $r(t) \geqslant r(t_1)$ for all $t > t_1$. By (5.19), $\dot{\omega}(t) \geqslant \zeta$ for some $\zeta > 0$ and all $t \geqslant t_1$ which is impossible.

Since $\omega^* > \pi/2$, for sufficiently large t we have $\omega(t) \in (\omega^* - \delta, \omega^*)$ with $\delta < \omega^* - (\pi/2)$. From (5.14) it follows that:

$$\dot{\omega}(t) \geqslant \cos^2(\omega^* - \delta) + o(r(t)), \tag{5.20}$$

which is impossible.

To prove the second part of the proposition observe that, if $r(t) \to 0$ for $t \to \infty$, we have

$$\dot{\omega}(t) \leqslant \cos^2 \omega(t) = \kappa(t)$$

with $\kappa(t) \to 0$ for $t \to \infty$. For $\omega \in ((\pi/2) + \kappa\pi - 1, (\pi/2) + \kappa\pi]$, κ integer, we have $\cos \omega \leqslant -(\omega - (\pi/2) - \kappa\pi)$, hence

$$\begin{aligned} \frac{d(\omega(\cdot) - (\pi/2) - k\pi)}{dt}(t) = \dot{\omega}(t) &\leqslant (w(t) - (\pi/2) - k\pi)^2 + \kappa(t) \\ &\leqslant ((\pi/2) + k\pi - \omega(t)) + \kappa(t). \end{aligned} \tag{5.21}$$

Let $k > 1$, and t_k be such that $w(t_k) = (\pi/2) + k\pi - 1$. Integrating (5.21) we obtain

$$\omega(t) - (\pi/2) - k\pi \leqslant -e^{-t-t_k} + \int_{t_k}^{t} e^{-(t-s)} \kappa(s)\, ds \tag{5.22}$$

while $\omega(t) \leqslant (\pi/2) + k\pi$. Let T_k be smallest positive such that $\omega(t_k + T_k) = \pi/2$. From (5.22) we obtian

$$e^{-T_k}(1 + \sup_{t \geqslant t_k} \kappa(t)) \leqslant \sup_{t \geqslant t_k} \kappa(t).$$

This proves $T_k \to \infty$ for $k \to \infty$ and completes the proof of the proposition. □

In order to turn the conclusion of this proposition into a statement about zeros of solutions of Eq. (1.1) note that, if z is a solution of the system (5.12), then the solution of Eq. (1.1) it represents satisfies

$$x(t) = z_1(t)\eta_0(0) + z_2(t)\sigma(0) + h(z(t))$$

with $h(z) = o(|z|)$ and $Dh(0) = 0$. Since $\eta_0(0) = 1, \sigma(0) = 0$, in the polar coordinates $z = (r\cos\omega, r\sin\omega)$ we have

$$x(t) = z_1(t) + h(|z(t)|) = r(t)\sin\omega(t) + \tilde{h}(r(t), \omega(t)), \tag{5.23}$$

where the function $(r, \omega) \mapsto \tilde{h}(r, \omega) = h(r\sin\omega, r\cos\omega)$ is C^1-smooth and satisfies $\tilde{h}(r, \omega) = o(r)$, $D_r\tilde{h}(0, \omega) = 0$ and

$$\begin{aligned} D_\omega\tilde{h}(r, \omega) &= r[D_{z_1}h(r\sin\omega, r\cos\omega)\cos\omega \\ &\quad - D_{z_2}h(r\sin\omega, r\cos\omega)\sin\omega] = o(r). \end{aligned} \tag{5.24}$$

We have

Corollary 5.2. *For $a = 1$ there is a neighbourhood V of 0 in C so that for every solution $x : [-1, \infty) \to \mathbb{R}$ of Eq. (1.1) with segments $0 \neq x_t \in W_c \cap V$ for all $t \geqslant 0$ the zeros in some ray $[t_x, \infty)$ form a strictly increasing sequence $(z_j)_0^\infty$ which tends to ∞. If in addition $\lim_{t\to\infty} x(t) = 0$ then $z_{j+1} - z_j \to \infty$ as $j \to \infty$.*

Proof. One has

$$|\sin\gamma| \geqslant 1 - |(2/\pi)(\gamma - (\pi/2) - k\pi)|$$

for k integer and $|\gamma - (\pi/2) - k\pi| \leqslant (\pi/2)$. Therefore, for any $\eta \in (0, 1)$, $\gamma \in \mathbb{R}$, and $k \in \mathbb{Z}$ with $|\gamma - (\pi/2) - k\pi| < (\pi/2)(1 - \eta)$ one has $|\sin\gamma| > \eta$ and, consequently, $|\sin\gamma - h(r, \gamma)/r| > 0$ for r sufficiently small. Therefore, for $r(t)$ sufficiently small all zeros of the function $t \mapsto r(t)\sin\omega(t) + \tilde{h}(r(t), \omega(t))$ are located in the intervals given by $|\omega(t) - k\pi| < \eta(\pi/2)$ for some integer k.

On the other hand, let $t_-^k < t_+^k$ be such that $\omega(t_\pm^k) = k\pi \pm \eta(\pi/2)$. Then, if $r(t)$ is sufficiently small, the function $t \mapsto x(t) = r(t)\sin\omega(t) + \tilde{h}(r(t), \omega(t))$ changes sign on $[t_-^k, t_+^k]$. Furthermore, using (5.13),(5.14) and (5.24) one obtains

$$\begin{aligned} \dot{x}(t) &= \dot{r}(t)\sin\omega(t) + r(t)\cos\omega(t)\dot{\omega}(t) \\ &\quad + D_r\tilde{h}(r(t), \omega(t))\dot{r}(t) + D_\omega\tilde{h}(r(t), \omega(t))\dot{\omega}(t) \\ &= r(t)\cos\omega(t)\sin^2\omega(t) + r(t)\cos\omega(t)\cos^2\omega(t) + o(r(t)) \\ &= r(t)\cos\omega(t) + o(r(t)). \end{aligned} \tag{5.25}$$

Hence, for $r(t)$ sufficiently small $\dot{x}(t)$ has the sign of $\cos \omega t$ on $[t_-^k, t_+^k]$ and, therefore, is monotonic. Consequently, it has a single zero in $[t_-^k, t_+^k]$.

Let now $r(t) \to 0$ for $t \to \infty$. Then, arguments as in the last part of the proof of Proposition 1 imply $t_-^{k+1} - t_+^k \to \infty$ for $k \to \infty$ and complete the proof. □

We have not been able to prove that $r(t)$ actually tends to 0 for $t \to \infty$. As some hint towards such a conclusion we show that this is the case for the truncated system of equations containing only the leading terms.

Proposition 5.2. *The zero equilibrium of the truncated system*

$$\begin{aligned} \dot{z}_1 &= z_2 - \tfrac{2}{3}|z_1|z_1, \\ \dot{z}_2 &= -2|z_1|z_1 \end{aligned} \tag{5.26}$$

is asymptotically stable.

Proof. We consider the (positive definite) C^{-1}-function $V : \mathbb{R}^2 \to \mathbb{R}$ given by

$$V(z_1, z_2) = |z_1|z_1^2 + \tfrac{3}{4}z_2^2. \tag{5.27}$$

By differentiating with respect to t along the solutions of (5.26) we obtain

$$\dot{V}(z_1, z_2) = 3|z_1|z_1\dot{z}_1 + \tfrac{3}{2}z_2\dot{z}_2 = 3|z_1|z_1z_2 - 2|z_1|^2{z_1}^2 - 3z_2|z_1|z_1 = -2z_1^4 \leqslant 0.$$

Thus V is a positive definite Lyapunov function for the system (5.26). We also have

$$E = \{(z_1, z_2) \in \mathbb{R}^2, |\dot{V}(z_1, z_2) = 0\} = \{(z_1, z_2) \in \mathbb{R}^2, |z_1 = 0\}.$$

It is obvious that the origin is the only invariant subset of E with respect to (5.26)

By [1, Theorem 5.5], the zero solution of (5.26) is asymptotically stable. □

Remark 5.3. In local bifurcation analysis one augments the vector of state variables by the parameter, in our case $\alpha = a - 1$. The resulting restriction to the (three-dimensional) center manifold at $z_1 = z_2 = \alpha = 0$ can be obtained in a similar way as (5.12) with q from (5.3) replaced by $q(\phi, \alpha) = -|\phi(0)|\phi(0) + \alpha(\phi(0) - \phi(1))$. One obtains

$$\begin{aligned} \dot{z}_1 &= z_2 - \frac{2}{3}(|z_1|z_1 + \alpha z_2) + o((|z_1| + |z_2| + |\alpha|)^2)\,, \\ \dot{z}_2 &= -2(|z_1|z_1 + \alpha z_2 + o((|z_1| + |z_2| + |\alpha|)^2)\,, \\ \dot{\alpha} &= 0. \end{aligned} \tag{5.28}$$

Using a C^1-transformation given by $u_1 = z_1, u_2 = z_2 - \frac{2}{3}(|z_1|z_1 + \alpha z_2) + o((|z_1| + |z_2| + |\alpha|)^2)$ we obtain the following system for u_1, u_2:

$$\begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= -2|u_1|u_1 - 2(\alpha u_2 - \tfrac{4}{3}|u_1|u_1) \\ &\quad + o((|u_1| + |u_2| + |\alpha|)^2). \end{aligned} \tag{5.29}$$

We see that for $\alpha = 0$ the restrictions of this sytem to the subsets $Z_\pm$ given by $0 \leqslant \pm z_1(u_1, u_2) = \pm u_1 + o((|u_1| + |u_2|)^2)$ coincide with the corresponding restrictions of the two versions of the generic Bogdanov–Takens form [13] glued together along the curve given by $z_1(u_1, u_2) = 0$. Unfortunately, because in the u_1, u_2-coordinates the sets $Z_\pm$ are not halfspaces and the conjugating homeomorphism may not preserve them, unlike in the true Bogdanov–Takens case, this does not mean that the system as a whole is topologically conjugate to the truncated one. The unfolding parameter α keeps the perturbed systems on the curve of the versal unfolding on which the systems have an equilibrium with a zero eigenvalue. It is a challenging problem to study the complete unfolding of the particular system (5.29) within the family of systems generated by delay equations.

6. CENTER-UNSTABLE MANIFOLDS

In this section we assume $a > 1$. A suitable modification of the proofs in Chapter IX in [5] shows that there exists a two-dimensional local center-unstable manifold W_{cu} of the semiflow F of Eq. (1.1) at the stationary point $0 \in C$. That is, W_{cu} is a C^1-submanifold of C with $0 \in W_{cu}$ and $T_0 W_{cu} = C_{0+} = L$ which is locally positively invariant under the semiflow. We may assume that the projection P along N onto L maps W_{cu} homeomorphically onto a open neighbourhood of 0 in L, and that the inverse map $I_{cu} : PW_{cu} \to C$ of the restriction $P|W_{cu}$ is continuously differentiable with injective derivatives.

The manifold W_{cu} contains a local center manifold W_c as in Section 5 and a local unstable manifold. The latter is a one-dimensional C^1-submanifold W_u of W_{cu} (or, of C) with $0 \in W_{cu}$ and $T_0 W_u = C_+ = \mathbb{R}\eta_u$ which is locally positively invariant under the semiflow and consists of segments $x_t, t \leqslant 0$, of solutions $x : \mathbb{R} \to \mathbb{R}$ of Eq. (1.1) which tend to 0 as $t \to -\infty$ (see, e.g., Chapter VIII in [5]).

The fact that the stationary point 0 is repelling in W_c means that there is a neighbourhood N_c of 0 in W_c so that for every $\phi \in N_c$ there is a solution $x^{(\phi)} : \mathbb{R} \to \mathbb{R}$ of Eq. (1.1) with $x_t^{(\phi)} \in W_c$ for all $t \leqslant 0$ and $x^{(\phi)}(t) \to 0$ as $t \to -\infty$. A consequence of this is that 0 is repelling also in

W_{cu}, in the sense that there exists a neighbourhood N_{cu} of 0 in $W_{cu} \cap C_{2a}$ with the following properties.

For each $\phi \in N_{cu} \setminus \{0\}$ there exists a solution $x^{(\phi)} : \mathbb{R} \to \mathbb{R}$ of Eq. (1.1) with $x_t^{(\phi)} \in W_{cu}$ for all $t \leqslant 0$, $x^{(\phi)}(t) \to 0$ as $t \to -\infty$, and $x_s^{(\phi)} \notin N_{cu}$ for some $s > 0$.

In case $0 \neq \phi \in N_{cu} \cap W_u$,

$$\operatorname{dist}(\|x_t^{(\phi)}\|^{-1} x_t^{(\phi)}, \{\eta_u, -\eta_u\}) \to 0 \quad \text{as } t \to -\infty.$$

In case $0 \neq \phi \in N_{cu} \setminus W_u$,

$$\operatorname{dist}(\|x_t^{(\phi)}\|^{-1} x_t^{(\phi)}, \{\eta_0, -\eta_0\}) \to 0 \quad \text{as } t \to -\infty.$$

Set

$$S_L^1 = \{\phi \in L : \|\phi\| = 1\}.$$

If $\phi \in N_{cu}, \psi \in N_{cu}, \phi \neq \psi$, then

$$\operatorname{dist}(\|x_t^{(\phi)} - x_t^{(\psi)}\|^{-1} (x_t^{(\phi)} - x_t^{(\psi)}), S_L^1) \to 0 \quad \text{as } t \to -\infty.$$

We consider the forward extension

$$W = F([0, \infty) \times N_{cu})$$

of N_{cu} and its closure $\bar{W}$.

Corollary 6.1. *We have $W \subset C_{2a}$. The closure $\bar{W}$ is compact. For every $\phi \in W (\phi \in \bar{W})$ there exists a solution $x^{(\phi)} : \mathbb{R} \to \mathbb{R}$ of Eq. (1.1) with $x_t^{(\phi)} \in W (x_t^{(\phi)} \in \bar{W})$ for all reals t.*

Proof. 1. Recall $N_{cu} \subset C_{2a}$. Proposition 2.3 gives $W \subset C_{2a}$. It follows that $\bar{W} \subset \overline{C_{2a}}$.

2 (Compactness). If $\phi \in W$ then $\phi = F(t, \chi)$ for some $t \geqslant 0$ and $\chi \in N_{cu}$. As $x^{(\chi)}(t) \to 0$ for $t \to -\infty$, there exist $\psi \in C_{2a}$ and $s \geqslant 1$ with

$$\phi = F(s, \psi) = F(1, F(s-1, \psi)) \in F(1, F(s-1, C_{2a})) \subset F(1, C_{2a}).$$

It follows that $W \subset F(1, C_{2a})$. Corollary 2.2 yields that $\bar{W}$ is compact.

3 (invariance of W). Consider $\phi \in W, \phi = F(t, \chi)$ with $t \geqslant 0$ and $\chi \in N_{cu}$. Then $\phi = x_t^{(\chi)}$. Consider the solution $x^{(\phi)} = x^{(\chi)}(t + \cdot)$ of Eq. (1.1). Let $s \in \mathbb{R}$ be given. There exists $r \leqslant 0$ so that $t + s + r \leqslant 0$ and $x_{t+s+r}^{(\chi)} \in W_{cu}$ so small that $x_{t+s+r}^{(\chi)} \in N_{cu}$. Hence

$$x_s^{(\phi)} = x_{(t+s)}^{(\chi)} = F(-r, x_{t+s+r}^{(\chi)}) \in W.$$

4 (invariance of $\bar{W}$). Let $\phi \in \bar{W}$, $\phi = \lim_{j\to\infty} \phi_j$, $\phi_j \in W$ for all integers $j \geqslant 1$. For every $t \geqslant 0$ we have $F(t,\phi) = \lim_{j\to\infty} F(t,\phi_j) \in \bar{W}$ since W is positively invariant under the semiflow. The existence of a solution $x^{(\phi)} : \mathbb{R} \to \mathbb{R}$ of Eq. (1.1) with $x_0^{(\phi)} = \phi$ and $x_t^{(\phi)} \in \bar{W}$ for all reals t follows provided we can show that for every integer $n < 0$ there exists $\phi_n \in \bar{W}$ with $F(-n, \phi_n) = \phi$. Proof of the last statement: Let an integer $n < 0$ be given. Consider the points $\phi_{nj} = x_n^{(\phi_j)}$, for $j \in \mathbb{N}$. Due to the compactness of $\bar{W}$ a subsequence of points $\phi_{nj_k}, k \in \mathbb{N}$, converges to some $\phi_n \in \bar{W}$ as $k \to \infty$. By continuity,

$$F(-n, \phi_n) = \lim_{k\to\infty} F(-n, \phi_{nj_k}) = \lim_{k\to\infty} \phi_{j_k} = \phi.$$

□

Corollary 6.2. *For every $\phi \in W \setminus \{0\}$ there exists $t = t(\phi) \in \mathbb{R}$ so that $x^{(\phi)}(s) \neq 0$ for all $s \leqslant t$.*

Proof. We have $\phi = F(t, \chi)$ for some $t \geqslant 0$ and $\chi \in N_{cu}$. Hence $x^{(\phi)} = x^{(\chi)}(t + \cdot)$. By $\phi \neq 0$, $\chi \neq 0$. We have

$$0 < e^{-u} = \min\ \eta_u < 1 = \min\ \eta_0.$$

Recall that for $s \to -\infty$,

$$\operatorname{dist}(\|x_s^{(\chi)}\|^{-1} x_s^{(\chi)}, \{-\eta_u, \eta_u\}) \to 0, \qquad \text{or}$$

$$\operatorname{dist}(\|x_s^{(\chi)}\|^{-1} x_s^{(\chi)}, \{-\eta_0, \eta_0\}) \to 0.$$

It follows that there exists $t_* \leqslant 0$ so that for all $s \leqslant t_*$ and for all $s' \in [-1, 0]$ we have

$$\frac{1}{2e^u} < ||x_s^{(\chi)}||^{-1} |x^{(\chi)}(s')|.$$

In particular,

$$0 \neq x^{(\chi)}(s) = x^{(\phi)}(t + s) \qquad \text{for all } s \leqslant t_*.$$

□

Proposition 6.1. *We have $\bar{W} - \bar{W} \subset Z \cup \{0\}$, the restricted projection $P|\bar{W}$ is injective, and its inverse $I : P\bar{W} \to C$ is continuous.*

Proof. 1. Proof of $W - W \subset Z \cup \{0\}$. Let $\phi_j \in W$, $j \in \{1,2\}$, with $\phi_1 \neq \phi_2$ be given. Then $\phi_j = F(t_j, \chi_j)$ with $t_j \geqslant 0$ and $\chi_j \in N_{cu}$ for $j \in \{1,2\}$. Using $x^{(\chi_j)}(t) \to 0$ as $t \to -\infty$ and $x_t^{(\chi_j)} \in W_{cu}$ for all $t \leqslant 0$ we find $t_0 \geqslant 0$ and $\psi_j \in N_{cu}$, $j \in \{1,2\}$, so that $\phi_j = F(t_0, \psi_j)$ for $j \in \{1,2\}$. We have $\psi_1 \neq \psi_2$. Since

$$\delta(s) = \|x_s^{(\psi_1)} - x_s^{(\psi_2)}\|^{-1}(x_s^{(\psi_1)} - x_s^{(\psi_2)}) \to S_L^1 \quad \text{as } s \to -\infty$$

and since S_L^1 is compact we find a sequence $(s_k)_0^\infty$ in $(-\infty, -5]$ with $s_k \to -\infty$ as $k \to \infty$ and a point $\sigma \in S_L^1$ so that

$$\delta(s_k) \to \sigma \qquad \text{for } k \to \infty.$$

We have $\sigma = \alpha\eta_0 + \beta\eta_u$ with $(\alpha, \beta) \in \mathbb{R}^2 \setminus \{(0,0)\}$, and the solution $v : [-1, \infty) \to \mathbb{R}$ of Eq. (4.1) with $v_0 = \sigma$ is given by

$$v(s) = \alpha + \beta e^{us} \quad \text{for all } s \geqslant -1.$$

We see that v is slowly oscillating and that there exists $s \in [0,2]$ so that $T_s\sigma = v_s$ has no zero. Choose $\epsilon > 0$ with

$$2\epsilon < \min_{-1 \leqslant s' \leqslant 0} |v_s(s')|.$$

By the continuity of $D_2F(s, \cdot)$ there exists $\rho > 0$ so that for all $\phi \in C_\rho$,

$$\|D_2F(s,\phi) - T_s\| \leqslant \epsilon.$$

Choose $k_0 \in \mathbb{N}$ so that for all integers $k \geqslant k_0$,

$$x_{s_k}^{(\psi_1)} \in C_\rho \qquad \text{and} \qquad x_{s_k}^{(\psi_2)} \in C_\rho.$$

For such k,

$$\begin{aligned}
&\|F(s, x_{s_k}^{(\psi_1)}) - F(s, x_{s_k}^{(\psi_2)}) - T_s(x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)})\| \\
&\quad = \| \int_0^1 D_2F(s, x_{s_k}^{(\psi_2)} + \theta(x_{s_k}^{(\psi_1)} \\
&\qquad - x_{s_k}^{(\psi_2)}))[x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}]d\theta - \int_0^1 T_s[x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}]\,d\theta \| \\
&\quad \leqslant \max_{\theta \in [0,1]} \|D_2F(s, x_{s_k}^{(\psi_2)} + \theta(x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)})) - T_s\| \|x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}\| \\
&\quad \leqslant \epsilon \|x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}\|.
\end{aligned}$$

Choose $k_1 \geqslant k_0$ so large that for all integers $k \geqslant k_1$,

$$\|T_s(\delta(s_k) - \sigma)\| < \epsilon.$$

For every integer $k \geqslant k_1$ we obtain

$$\begin{aligned}
&\| \|x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}\|^{-1}(F(s, x_{s_k}^{(\psi_1)}) - F(s, x_{s_k}^{(\psi_2)})) - T_s\sigma\| \\
&\quad \leqslant \| \|x_{s_k}^{(\psi_1)} - x_{s_k}^{(\psi_2)}\|^{-1}(F(s, x_{s_k}^{(\psi_1)}) - F(s, x_{s_k}^{(\psi_2)})) \\
&\qquad - T_s\delta(s_k)\| + \|T_s\delta(s_k) - T_s\sigma\| \\
&\quad \leqslant \epsilon + \epsilon = 2\epsilon.
\end{aligned}$$

It follows that for such k the function:

$$F(s, x_{s_k}^{(\psi_1)}) - F(s, x_{s_k}^{(\psi_2)})) = x_{s+s_k}^{(\psi_1)} - x_{s+s_k}^{(\psi_2)}$$

has no zero. We apply Proposition 3.1 (i) to the solution

$$y = x^{(\psi_1)}(s + s_k + \cdot) - x^{(\psi_2)}(s + s_k + \cdot)$$

of Eq. (3.1) with

$$h(t, \eta) = f(\eta + x^{(\psi_2)})(s + s_k + t)) - f(x^{(\psi_2)}(s + s_k + t))$$

and find

$$Z \ni y_{t_0-(s+s_k)} = \phi_1 - \phi_2.$$

2. Proof of $\bar{W} - \bar{W} \subset Z \cup \{0\}$. Let $\phi_j \in \bar{W}$, $j \in \{1, 2\}$, with $\phi_1 \neq \phi_2$ be given. By Corollary 1,

$$\psi_j = x_{-3}^{(\phi_j)} \in \bar{W} \quad \text{for } j \in \{1, 2\}$$

and $\psi_1 \neq \psi_2$. Using part 1 of the proof we find

$$0 \neq \psi_1 - \psi_2 \in \bar{W} - \bar{W} \subset \overline{Z \cup \{0\}} \subset \overline{S_1 \cup \{0\}} = \overline{S}_1 = S_1.$$

Proposition 3.1(i) yields

$$\phi_1 - \phi_2 = F(3, \psi_1) - F(3, \psi_2) \in Z.$$

3. The restriction $P|\bar{W}$ is injective since for all ϕ_1, ϕ_2 in $\bar{W}$ with $\phi_1 \neq \phi_2$ we have

$$\phi_1 - \phi_2 \in Z \subset C \setminus N = C \setminus P^{-1}(0)$$

or $0 \neq P(\phi_1 - \phi_2) = P\phi_1 - P\phi_2$.

4. The compactness of $\bar{W}$ yields that the inverse $I : P\bar{W} \to C$ of the continuous injective map $P|\bar{W}$ is continuous. □

Corollary 6.3. *The set PW is open in L, and $\overline{PW} = P\bar{W}$.*

Proof. 1 (Openness). Let $\chi \in PW$ be given. Then $\chi = P\phi$ with $\phi \in W$, and $\phi = F(t, \psi)$ with $t \geqslant 0$ and $\psi \in N_{cu} \subset W_{cu}$. The composition $P \circ F(t, \cdot) \circ I_{cu}$ has range in $PW \subset L$ and has injective derivatives, due to Propositions 2.2 and 6.1. It follows that the set $(P \circ F(t, \cdot) \circ I_{cu})(N_{cu}) \subset PW$ contains an open neighbourhood of $\chi = PF(t, I_{cu}(P\psi))$ in L.

2. The continuity of P yields $P\bar{W} \subset \overline{PW}$. The inclusion $PW \subset P\bar{W}$ and the compactness of $P\bar{W}$ combined imply $\overline{PW} \subset P\bar{W}$. □

Corollary 6.4. *For every $\phi \in \bar{W} \setminus \{0\}$ the solution $x^{(\phi)}$ of Eq. (1.1) is slowly oscillating, and all of its zeros are simple.*

Proof. The fact that $x = x^{(\phi)}$ is slowly oscillating is obvious from

$$0 \neq x_t = x_t - 0 \in \bar{W} - \bar{W} \subset Z \cup \{0\}.$$

Simplicity of any zero z of x follows from:

$$\dot{x}(z) = 0 - ax(z-1) \neq 0.$$

□

The preceding result does not imply that $x^{(\phi)}$ actually has zeros. The proof that the latter is indeed the case employs the facts that the stationary point is repelling in local center and center-unstable manifolds and that the projection P is injective on the forward extension W and on its closure.

Corollary 6.5. *Let $\phi \in \bar{W} \setminus \{0\}$ be given. Then every interval $[t, \infty), t \in \mathbb{R}$, contains a zero of $x^{(\phi)}$. In case $0 \neq \phi \in W$ the Zeros of $x^{(\phi)}$ form a strictly increasing sequence $(Z_j(\phi))_0^\infty$ .In case $0 \neq \phi \in \bar{W} \setminus W$ the zeros of $x^{(\phi)}$ form a strictly increasing sequence $(Z_j(\phi))_{-\infty}^{\infty}$. In both cases,*

$$z_j(\phi) + 1 < z_{j+1}(\phi).$$

Proof. 1. Let $0 \neq \phi \in \bar{W}, x = x^{(\phi)}, t \in \mathbb{R}$. Suppose x has no Zero in $[t, \infty)$. Then $x_s \to 0$ for $s \to \infty$, due to Proposition 3.2. Consider $N_{cu} \subset \bar{W}$. There exists $s_0 \in \mathbb{R}$ so that for all real $s \geqslant s_0$ the projected segment P_{x_s} is contained in the neighbourhood PN_{cu} of 0 in L. Hence $x_s \in N_{cu}$ for all $s \geqslant s_0$,in contradiction to the fact that flowlines starting at points in $N_{cu} \setminus \{o\}$ must leave N_{cu}.

2. In case $0 \neq \phi \in W$ we have from Corollary 6.2 in combination with part 1 that there is a smallest zero $z_0 = z_0(\phi) of\, x = x^{(\phi)}$. Proposition 3.1 (ii) gives $\operatorname{sign}(x(t)) = -\operatorname{sign}(x(z_0 - 1)) \neq 0$ for $z_0 < t \leqslant Z_0 + 1$. Define $z_1 = z_1(\phi)$ to be the smallest zero of x in $(z_0 + 1, \infty)$. Proceed by induction.

3. Let $\phi \in \bar{W} \setminus W$ be given. Set $x = x^{(\phi)}$. By Corollary 6.1, $x_t \in \bar{W} \setminus W$ for all reals t. The assertion about the sequence of zeros follows easily by means of Corollary 6.4 and the result of part 1 provided the set of zeros of x is not bounded from below. Suppose the last statement is false, i.e., there exists $t_0 \in \mathbb{R}$ with $x(t) \neq 0$ for all reals $t \leqslant t_0$. Consider the case $0 < x(t)$ for all $t < t_0$. As $x_t \in \bar{W} \setminus W \subset \bar{W} \setminus N_{cu}$ for all $t \in \mathbb{R}$ we have that P_{x_t} avoids the neighbourhood PN_{cu} of 0 in L. It follows that there exists $c > 0$ with $c \leqslant \|x_t\|$ for all $t \in \mathbb{R}$. Hence the α-limit set $\alpha(x) \subset \bar{W}$ consists of nonnegative functions $\psi \in C$ with $||\psi|| \geqslant c$. The solutions $y : \mathbb{R} \to \mathbb{R}$ of Eq.(1.1) with segments in $\alpha(x)$ are slowly oscillating, nonnegative, and do not converge to 0 as $t \to \infty$. This yields a condradiction to Proposition 3.2.
The argument in case $x(t) < 0$ for $t \leqslant t_0$ is analogous. □

The oscillatory behavior of the solutions $x^{(\phi)} : \mathbb{R} \to \mathbb{R}, 0 \neq \phi \in \bar{W}$, implies that the projected flowlines $\mathbb{R} \ni t \mapsto P_{x_t}^{(\phi)} \in L$ wind around the origin in L. In terms of the basis $\{\phi_0, \phi_-\}$ chosen in Section 4 this winding motion is easily described as follows.

Corollary 6.6. *Let $0 \neq \phi \in \bar{W}, x = x^{(\phi)}, z_j = z_j(\phi)$ for some integer j. In case $0 < \dot{x}(z_j)$ we have*

$$Px_{z_j} \in (-\infty, 0)\phi_0,$$

$$Px_t \in (-\infty, 0)\phi_0 + (0, \infty)\phi_- \quad for \quad z_j < t < z_j + 1,$$

$$Px_{z_j+1} \in (0, \infty), \phi_-,$$

$$Px_t \in (0, \infty)\phi_0 + (0, \infty)\phi_- \quad for \quad z_j + 1 < t < z_j + 1,$$

$$Px_{z_j+1} \in (0, \infty), \phi_0.$$

If $0 \neq \phi \in W$, $j = 0$, and $t < z_0$ then

$$Px_t \in (-\infty, 0)\phi_0 + (-\infty, 0)\phi_-.$$

Proof. Use Proposition 4.3. □

Of course, there is an analogue of Corollary 6.6 for the case $\dot{x}(z_j) < 0$.

The corollaries 6.4 and 6.5 combined imply that for every $\phi \in \bar{W} \setminus \{0\}$ in the hyperplane H_0 the slowly oscillating solution $x^{(\phi)} : \mathbb{R} \to \mathbb{R}$ has a smallest zero $\zeta_1(\phi)$ in $(0, \infty)$, with $\zeta_1(\phi) > 1$, and a smallest zero $\zeta_2(\phi)$ in $(\zeta_1(\phi), \infty)$; we have $\zeta_2(\phi) > \zeta_1(\phi) + 1$.

Proposition 6.2. *The map*

$$(\bar{W} \setminus \{0\} \cap H_0 \ni \phi \mapsto \zeta_2(\phi) \in \mathbb{R}$$

is continuous.

Proof. Use continous dependence on initial data, simplicity of the zeros $\zeta_1(\phi)$ and $\zeta_2(\phi)$, and the fact that on bounded subintervals of $[1, \infty)$ also derivatives of solutions $x : [-1, \infty) \to \mathbb{R}$ depend continuously on initial data. □

7. A PERIODIC ORBIT

In this section we assume $a > 1$ and prove the following result.

Theorem 7.1. *The set $\bar{W} \setminus W$ is the orbit $\mathcal{O} = \{p_t : t \in \mathbb{R}\}$ of a slowly oscillating periodic solution $p : \mathbb{R} \to \mathbb{R}$ of Eq.(1.1),whose minimal period is given by 3 consecutive zeros. We have*

$$\operatorname{int}(P\mathcal{O}) = PW$$

and for every $\phi \in W \setminus \{0\}$,

$$\operatorname{dist}(F(t, \phi), \mathcal{O}) \to 0 \quad as \ t \to \infty.$$

Proof. 1 (Construction of the periodic orbit).

1.1 Consider the semi-axes $(0, \infty)\phi_0$ in L. As PW is a bounded open neighborhood of 0 in L we have that

$$\rho = \inf\{r > 0 : r\phi_0 \notin PW\}$$

is finite and positive. For $0 < r < \rho, r\phi_0 \in PW$. Also $\rho\phi_0 \in \overline{PW} = P\bar{W}$. From the openness of PW in L one easily concludes that $\rho\phi_0 \notin PW$. Let $\phi = I(\rho\phi_0) \in \bar{W} \setminus W$ and define $p = x^{(\phi)}$.Then $p_t \in \bar{W} \setminus W$ for all $t \in \mathbb{R}$. For each integer j, set $z_j = z_j(\phi)$.We have $p_0 = \phi \in H_0$(see Proposition 4.4),or $p(0) = 0$.So we may assume $z_0 = 0$. From $\rho\phi_0 = P\phi = \phi(-1)\phi_0$ we infer $\phi(-1) > 0$. It follows that:

$$Pp_t \in (0, \infty)\phi_0 + (-\infty, 0)\phi_- \quad \text{for } 0 < t < 1$$

and

$$Pp_1 = p(1)\phi_- \in (-\infty, 0)\phi_-.$$

Notice that $Pp_t \notin PW$.Our next aim is to prove

$$(p(1), 0)\phi_- \subset PW.$$

This requires a bit of preparation.

1.2 The line segment $[0, \rho]\phi_0$, the arc $\{P_{p_t} : 0 < t < 1\}$, and the line segment $[p(1), 0]\phi_-$ form the trace of a simple closed curve c. All points in the complement

$$L \setminus ([0, \infty)\phi_0 + (-\infty, 0]\phi_-)$$

of the closed sector which contains the trace $|c|$ can be connected by rays in $L \setminus |c|$ to points with arbitrarily large norm, and must therefore belong to the set ext (c). By the same argument,

$$(\rho, \infty)\phi_0 \subset \mathrm{ext}(c) \quad \text{and} \quad (-\infty, p(1))\phi_- \subset \mathrm{ext}(c).$$

Proof that for every $r \in (p(1), 0)$ there exists $\epsilon = \epsilon(r) > 0$ with

$$(0, \epsilon)\phi_0 + (r - \epsilon, r + \epsilon)\phi_- \subset \mathrm{int}(c).$$

Choose $\epsilon > 0$ with

$$((-\epsilon, \epsilon)\phi_0 + (r - \epsilon, r + \epsilon)\phi_-) \cap |c| = (r - \epsilon, r + \epsilon)\phi_-.$$

From the preceeding statements involving ext(c) we have

$$(-\epsilon, 0)\phi_0 + (r - \epsilon, r + \epsilon)\phi_- \subset \mathrm{ext}(c).$$

The set $(0, \epsilon)\phi_0 + (r - \epsilon, r + \epsilon)\phi_-$ is open, connected, and disjoint with $|c|$. Therefore, it is either contained in $\mathrm{int}(c)$ or contained in $\mathrm{ext}(c)$. In the last case, $r\phi_- \in |c|$ is not in the closure of $\mathrm{int}(c)$, which contradicts $|c| = \partial\,\mathrm{int}(c)$.

1.3 Proof of $(p(1), 0)\phi_- \subset PW$. Suppose $r\phi_- \notin PW$ for some $r \in (p(1), 0)$. Set

$$\rho_* = \sup\{s < 0 : s\phi_- \notin PW\}.$$

Arguing as in part 1.1 we find $r \leqslant \rho_* < 0$ and $\rho_*\phi_- \in P\bar{W} \setminus PW$. Set $\phi_* = I(\rho_*\phi_-)$, $x^* = x^{(\phi_*)}$, and $x = x^*(\cdot - 1)$.From

$$\rho_*\phi_- = Px_0^* = x^*(-1)\phi_0 + x^*(0)\phi_- = x(0)\phi_0 + x(1)\phi_-$$

we infer $x(0) = 0$ and $x(1) < 0$. For $0 < t < 1$,

$$Px_t \in (0, \infty)\phi_0 + (-\infty, 0)\phi_-.$$

Let $\epsilon_* = \epsilon(p_*)$.Using $P_{x_1} = P_{x_0^*} = \rho_*\phi_-$ and continuity we obtain that for $t < 1$ sufficiently close to 1,

$$Px_t \in (-\epsilon_*, \epsilon_*)\phi_0 + (\rho_* - \epsilon_*, \rho_* + \epsilon_*)\phi_-.$$

Using also the last statement about the position of Px_t for all $t \in (0, 1)$ we get

$$Px_t \in (0, \epsilon_*)\phi_0 + (\rho_* - \epsilon_*, \rho_* + \epsilon_*)\phi_- \subset \mathrm{int}(c)$$

for all $t < 1$ sufficiently close to 1. Now consider

$$Px_0 = x(-1)\phi_0 + x(0)\phi_- = x(-1)\phi_0.$$

As $x_0 \in \bar{W} \setminus W$ we have $Px_0 \notin PW$. The definition of ρ yields $x(-1) \geqslant \rho$. In case $x(-1) = \rho$ we get $Px_0 = \rho\phi_0 = Pp_0$, hence $x_0 = p_0$, consequently $x_1 = p_1$, in contradiction to

$$Px_1 = Px_0^* = \rho_*\phi_- \neq p(1)\phi_- = Pp_1.$$

In case $x(-1) > \rho$ we have $Px_0 \in \mathrm{ext}(c)$. As $Px_t \in \mathrm{int}(c)$ for some $t \in (0, 1)$ we conclude that there exists $s \in (0, 1)$ with $Px_s \in |c|$. Since Px_s is contained in $(0, \infty)\phi_0 + (-\infty, 0)\phi_-$ we have

$$Px_s \notin [0, \rho]\phi_0 \cup [p(1), 0]\phi_-.$$

Using $Px_s \in |c|$ we infer that for some $s' \in (0, 1)$ we have $Px_s = Pp_{s'}$. In case $s = s'$ we get $x_1 = p_1$, which yields a contradiction as above. In case $s < s'$ we get

$$F(s', x_0) = F(s, F(s' - s, p_0)) = F(s, p_{s'-s}).$$

As $Px_0 \in (0, \infty)\phi_0$ and $Pp_{s'-s} \in (0, \infty)\phi_0 + (-\infty, 0)\phi_-$ we have $x_0 \neq p_{s'-s}$, and the last equation contradicts Proposition 2.2. In case $s' < s$ we obtain a contradiction in the same way.

1.4 Consider

$$Pp_{z_1} = p(z_1 - 1)\phi_0 \in (-\infty, 0)\phi_0,\ Pp_{z_1+1} = p(z_1 + 1)\phi_- \in (0, \infty)\phi_-$$

and $Pp_{z_2} = p(z_2 - 1)\phi_0 \in (0, \infty)\phi_0$. Arguing as before we find

$$(p(z_1 - 1), 0)\phi_0 \subset PW,\ (0, p(z_1 + 1))\phi_- \subset PW, \quad \text{and } (0, p(z_2 - 1))\phi_0 \subset PW.$$

The last inclusion, the relation $Pp_{z_2} = p(z_2 - 1)\phi_0 \notin PW$, and the definition of ρ combined yield $p(z_2 - 1) = \rho$, hence $Pp_{z_2} = Pp_0$, and consequently $p_{z_2} = p_0$. It follows that p is periodic with minimal period $z_2 = z_2 - 0 = z_2 - z_0$.

Notice that the curve $\mathbb{R} \ni t \mapsto p_t \in C$ is continuously differentiable. Set $\mathcal{O} = \{p_t : t \in \mathbb{R}\} = \{p_t : 0 \leqslant t \leqslant z_2\}$.

2. Proof of $PW \subset \mathrm{int}(P\mathcal{O})$. The set PW is open (Corollary 6.3) and connected since it contains a neighbourhood of 0 and $PW \ni Px_t^{(\phi)} \to 0$ as

$t \to -\infty$ for all $\phi \in W$. Also, $PW \cap P\mathcal{O} = \emptyset$ since $\mathcal{O} \subset \bar{W} \setminus W$. It follows that $PW \subset \mathrm{int}(P\mathcal{O})$ or $PW \subset \mathrm{ext}(P\mathcal{O})$. The simple closed curve

$$c : [0, z_2] \ni t \mapsto Pp_t \in L$$

is continuously differentiable, with $c(t) = p(t-1)\phi_0 + p(t)\phi_-$. Hence $c'(0) = p'(-1)\phi_0 + p'(0)\phi_-$. As all zeros of p are simple, $p'(0) \neq 0$, and we infer that $c'(0)$ and ϕ_0 are linearly independent. We have $|P\mathcal{O}| \cap (0, \infty)\phi_0 = \{p(-1)\phi_0\}$. It follows that the unbounded ray $(p(-1), \infty)\phi_0$ is contained in $\mathrm{ext}(P\mathcal{O})$. Using a remark in Section 1 for a suitable reparametrization of $P\mathcal{O}$ we infer

$$(p(-1) - \epsilon, p(-1))\phi_0 \subset \mathrm{int}(P\mathcal{O})$$

for some $\epsilon > 0$. As $(0, p(-1))\phi_0 \subset PW$ we find

$$PW \subset \mathrm{int}(P\mathcal{O}).$$

3. Proof of $\mathrm{int}(P\mathcal{O}) \subset PW$. Suppose there exists $\chi \in \mathrm{int}(P\mathcal{O}) \setminus PW$. There is a curve $c : [0, 1] \to L$ in $\mathrm{int}(P\mathcal{O})$ with endpoints $c(0) = 0 \in PW \subset \mathrm{int}(P\mathcal{O})$ and $c(1) = \chi$. As PW is open we have that

$$\bar{s} = \inf\{s \in [0, 1] : c(s) \notin PW\}$$

satisfies $0 < \bar{s} \leqslant 1$, $c([0, \bar{s})) \subset PW$, and

$$c(\bar{s}) \in \overline{PW} \setminus PW = P\bar{W} \setminus PW = P(\bar{W} \setminus W).$$

In particular, $c(\bar{s}) \neq 0$. Set $\psi = I(c(\bar{s})) \in \bar{W} \setminus W$ and $x = x^{(\psi)}$. Then $\psi \notin \mathcal{O}$ (since $P\psi = c(\bar{s}) \in \mathrm{int}(P\mathcal{O})$). Hence $x_t \notin \mathcal{O}$ for all reals t. Consequently, $Px_t \notin P\mathcal{O}$ for all reals t. There is a zero $z > 0$ of x so that $Px_z \in (0, \infty)\phi_0$. Recall $P\mathcal{O} \cap (0, \infty)\phi_0 = \{p(-1)\phi_0\}$ and $(p(-1), \infty)\phi_0 \subset \mathrm{ext}(P\mathcal{O})$. It follows that $Px_z \in (0, p(-1))\phi_0$ or $Px_z \in (p(-1), \infty)\phi_0$. In the first case, $Px_z \in P\bar{W} \setminus PW$ yields a contradiction to $(0, p(-1))\phi_0 \subset PW$. In the second case the restriction $c|[0, \bar{s}]$ and the map $[0, z] \ni t \mapsto Px_t \in L$ define a curve in $L \setminus (P\mathcal{O})$ with endpoints $0 \in \mathrm{int}(P\mathcal{O})$ and $Px_z \in \mathrm{ext}(P\mathcal{O})$, which is impossible.

4. Proof of $\bar{W} \setminus W = \mathcal{O}$. Recall $\mathcal{O} \subset \bar{W} \setminus W$. We have

$$\begin{aligned} P\bar{W} &= \overline{PW} = \overline{\mathrm{int}(P\mathcal{O})} \\ &= \mathrm{int}(P\mathcal{O}) \cup \partial(\mathrm{int}(P\mathcal{O})) \\ &= \mathrm{int}(P\mathcal{O}) \cup P\mathcal{O} = PW \cup P\mathcal{O} = P(W \cup \mathcal{O}) \subset P\bar{W}, \end{aligned}$$

hence $P\bar{W} = P(W \cup \mathcal{O})$, and thereby $\bar{W} = W \cup \mathcal{O}$.

5. Let $\phi \in W \setminus \{0\}$ be given. It remains to show that $\mathrm{dist}(F(t, \phi), \mathcal{O}) \to 0$ as $t \to \infty$. Set $x = x^{(\phi)}$ and $z_j = z_j(\phi)$ for all integers $j \geqslant 0$. As $x_t \to 0$ for

$t \to -\infty$ the curve $\mathbb{R} \ni t \mapsto x_t \in W \setminus \{0\}$ is not periodic, hence injective. Proposition 6.1 yields that the projected curve $\mathbb{R} \ni t \mapsto Px_t \in PW \setminus \{0\}$ is injective. We have $\dot{x}(z_0) \neq 0$ (Corollary 6.4). Consider the case $\dot{x}(z_0) > 0$. Then $Px_{z_0} = x(z_0 - 1)\phi_0, r = x(z_0 - 1) < 0$, and $Px_{z_{2j}} = x(z_{2j} - 1)\phi_0$ for all integers $j > 0$.

5.1. Proof of $x(z_2 - 1) < x(z_0 - 1)$. The set $\{P_{x_t} : t < z_0\} \cup [r, 0]\phi_0$ is the trace of a simple closed curve c in the closed sector

$$(-\infty, 0]\phi_0 + (-\infty, 0]\phi_- \subset L$$

with

$$Px_t = x(t-1)\phi_0 + x(t)\phi_- \in (-\infty, 0)\phi_0 + (-\infty, 0)\phi_- \text{for } t < z_0$$

(see Corollary 6.6). We have

$$\operatorname{int}(c) \subset (-\infty, 0)\phi_0 + (-\infty, 0)\phi_-$$

because each point in $L \setminus (|c| \cup ((-\infty, 0)\phi_0 + (-\infty, 0)\phi_-))$ can be connected by a ray in $L \setminus |c|$ to points with arbitrarily large norm, and must therefore belong to $\operatorname{ext}(c)$. As in part 1.2 above one sees that for each $\bar{r} \in (r, 0)$ there exists $\bar{\epsilon} = \epsilon(\bar{r}) > 0$ so that

$$\bar{r}\phi_0 + (-\bar{\epsilon}, \bar{\epsilon})\phi_0 + (-\bar{\epsilon}, 0)\phi_- \subset \operatorname{int}(c). \tag{7.1}$$

Set $\bar{r} = x(z_2 - 1) < 0$. The case $\bar{r} = r$ is impossible since $t \mapsto Px_t$ is injective. Assume $r < \bar{r} < 0$. Set $\bar{\epsilon} = \epsilon(\bar{r})$, so that (1) holds. We have

$$Px_{z_1+1} = x(z_1 + 1)\phi_- \in (-\infty, 0)\phi_- \subset \operatorname{ext}(c).$$

The curve $[z_1 + 1, z_2] \ni t \mapsto Px_t \in L$ has endpoints Px_{z_1+1} and Px_{z_2} and satisfies

$$Px_t \in (-\infty, 0)\phi_0 + (-\infty, 0)\phi_- \quad \text{for } z_1 + 1 < t < z_2. \tag{7.2}$$

The open neighbourhood

$$\bar{r}\phi_0 + (-\bar{\epsilon}, \bar{\epsilon})\phi_0 + (-\bar{\epsilon}, \bar{\epsilon})\phi_-$$

of $\bar{r}\phi_0 = Px_{z_2}$ in L contains points Px_t with $z_1 + 1 < t < z_2$, due to continuity. Using (7.2) and (7.1) we infer

$$Px_t \in \bar{r}\phi_0 + (-\bar{\epsilon}, \bar{\epsilon})\phi_0 + (-\bar{\epsilon}, 0)\phi_- \subset \operatorname{int}(c)$$

for some $t \in (z_1 + 1, z_2)$. Using also $Px_{z_1+1} \in \operatorname{ext}(c)$ we find $s \in (z_1 + 1, t)$ with

$$Px_s \in |c| \cap ((-\infty, 0)\phi_0 + (-\infty, 0)\phi_-) = \{Px_t : t < z_0\}$$

in contradiction to the injectivity of $t \mapsto Px_t$. Therefore, $\bar{r} < r$, or $x(z_2 - 1) < x(z_0 - 1)$.

5.2 Using similar arguments and induction one finds

$$x(z_{2j+2} - 1) < x(z_{2j} - 1) < 0 \quad \text{for all integers } j \geqslant 0.$$

As x is bounded the decreasing sequence $(x(z_{2j} - 1))_0^\infty$ converges to some $\xi < 0$. We have

$$\xi\phi_0 = \lim_{j\to\infty} x(z_{2j} - 1)\phi_0 = \lim_{j\to\infty} Px_{z_{2j}} \in \overline{PW} = P\bar{W}.$$

Proposition 4.4 guarantees that

$$\psi = I(\xi\phi_0) \in \bar{W}$$

belongs to H_0, i.e., $\psi(0) = 0$. By $\xi < 0$, $\psi \neq 0$. As I is continuous,

$$x_{z_{2j}} = I(Px_{z_{2j}}) \to I(\xi\phi_0) = \psi \quad \text{for } j \to \infty.$$

By Proposition 6.2,

$$z_{2j+2} - z_{2j} = \zeta_2(x_{z_{2j}}) \to \zeta_2(\psi) \quad \text{as } j \to \infty.$$

Consequently,

$$\psi = \lim_{j\to\infty} x_{z_{2j+2}} = \lim_{j\to\infty} F(z_{2j+2} - z_{2j}, x_{z_{2j}}) = F(\zeta_2(\psi), \psi).$$

It follows that $y = x^{(\psi)}$ is a periodic solution of Eq. (1.1), with $y_t \in \bar{W}\setminus\{0\}$ for all $t \in \mathbb{R}$. As solutions with segments in W tend to 0 as $t \to -\infty$ we have $y_t \in \bar{W}\setminus W$ for all $t \in \mathbb{R}$. In particular, $\psi \in \bar{W}\setminus W$. It follows that

$$P\psi \in P(\bar{W}\setminus W) = P\bar{W}\setminus PW = \overline{PW}\setminus PW = P\mathcal{O} \text{ (see part 4).}$$

Hence $\psi \in \mathcal{O}$. Consequently, $y = p(s + \cdot)$ for some $s \in \mathbb{R}$, and $\mathcal{O} = \{y_t : t \in \mathbb{R}\}$.

5.3. Proof of $\operatorname{dist}(x_t, \mathcal{O}) \to 0$ as $t \to \infty$. Let $\epsilon > 0$ be given. Using Proposition 6.2 and continuous dependence on initial data we find a neighbourhood N_ϵ of ψ in C such that for all $\chi \in N_\epsilon \cap \bar{W} \cap H_0$ we have

$$\zeta_2(\chi) \leqslant \zeta_2(\psi) + 1$$

and

$$||F(t, \chi) - F(t, \psi)|| < \epsilon \quad \text{for } 0 \leqslant t \leqslant \zeta_2(\psi) + 1.$$

There exists an integer $j = j(\epsilon) \geqslant 0$ so that for all integers $k \geqslant j, x_{z_{2k}} \in N_\epsilon$. Let $t \geqslant z_{2j}$. Then $z_{2k} \leqslant t < z_{2k+2}$ for some integer $k \geqslant j$. It follows that:

$$\begin{aligned} \operatorname{dist}(x_t, \mathcal{O}) &\leqslant ||x_t - F(t - z_{2k}, \psi)|| \text{ (since } F(t - z_{2k}, \psi) \in \mathcal{O}) \\ &= ||F(t - z_{2k}, x_{z_{2k}}) - F(t - z_{2k}, \psi)|| < \epsilon \end{aligned}$$

since

$$0 \leqslant t - z_{2k} < z_{2k+2} - z_{2k} = \zeta_2(x_{z_{2k}}) \leqslant \zeta_2(\psi) + 1.$$

□

8. CONCLUDING REMARKS

In Section 1 we summarized the messages the investigation of Eq. (1.1) sends to economic theory as well as the mathematics of the study itself. None of it is fully satisfactory and requires further development.

On the side of application: Although the model exhibits permanent fluctuations, it falls short of explaining their erratic nature observed in practice. Even worse, the analysis of the equation indicates that the essential long term dynamics takes place on a disk and, thus, is subject to the Poincaré –Bendixson theory which says that all recurrent motions are periodic. Therefore, new and more complex models are probably needed to produce more complicated dynamics.

For a full justification of the numerical simulations and, thus, also of the implications to economic theory a proof of global stability of the equilibrium 0 as well as a proof of generic convergence of solutions to the periodic orbit would be needed.

Mathematically, the dominant parts of the paper are the proof of existence of the periodic orbit and the auxiliary results. The idea to find the periodic orbit as the boundary of a two-dimensional unstable manifold of the equilibrium goes back to earlier work [16, 17, 19]. However, there are new aspects which seem worth to be mentioned.

First of all, in [16, 17] – as well as in other work on periodic solutions of delay differential equations (see [5] for reference) – the crucial fact that slowly oscillating solutions are actually oscillating, i.e., have infinitely many zeros on the positive semi-axis, is easily established by elementary arguments, for the parameters and initial data in question. The situation studied here is different and more difficult: In Section 3 elementary arguments led us only to the conclusion that slowly oscillating solutions with a bounded set of zeros are restricted by a necessary condition (Proposition 3.2; such solutions decay to 0 as $t \to \infty$). The proof in Section 6 that there exist infinitely many positive zeros works only for solutions in the compact global center-unstable manifold $\overline{W}$. It replies on the somewhat unexpected

fact (see Remark 5.1) that for $a > 1$ the equlibrium is repelling in the center manifold, and on the graph representation of $\bar{W}$.

A second major difference to former approaches lies in the choice of the co-ordinate system for the graph representation of $\bar{W}$. Whereas in [16, 17, 19] the decomposition of the state space C into the analogue of our space L and into the complementary generalized eigenspace Q is used for a graph representation of an unstable manifold, we choose here another complementary space N. The advantage of this choice is that for trajectories $t \mapsto x_t$ of slowly oscillating solutions $x : t \mapsto \mathbb{R}$ the spiraling motion of projections along N into the plane L is almost obvious, with the axes of a co-ordinate system in L as global transversals for the projected trajectories. In contrast to this, the description of the projected trajectories in [16, 17], in particular close to the boundary, requires some effort and uses, among others, a-priori estimates which express that slowly oscillating solutions do not decay to zero faster than exponentials. In the present work we can avoid such a-priori estimates. Let us add, however, that those a-priori estimates yield that the invariant manifold considered is a Lipschitz graph.

The choice of the complementary space N in section 4 is inspired by a choice of a co-ordinate system in [12].

The new co-ordinate system also sheds light on the Poincaré–Bendixson type results for delay differential equations which were obtained by Kaplan and Yorke [10, 11] and Mallet–Paret and Sell [14]. Their proofs associate curves $t \mapsto (x(t), x(t-1))$ or $t \mapsto (x(t), \dot{x}(t))$ in the plane $\mathbb{R}^2$ to certain solutions x, instead of flowlines $t \mapsto x_t$ in the space C of initial data. It is now clear what is behind the nice behavior of such curves in the plane, at least in case of our Eq. (1.1) and for similar equations: The curves in the plane are obtained from the flowlines $t \mapsto x_t$ in the compact invariant set $\bar{W}$ by the global chart on $\bar{W}$ which projects along N into L and then assigns to the projected point its coefficients with respect to the basis $\{\phi_{-1}, \phi_0\}$ of the space L.

Accordingly proofs of uniqueness of periodic orbits due to Nussbaum [15], Cao [3], and others, which employ planar curves as above, can now be reinterpreted in terms of behavior of flowlines in the state space C.

We mentioned problems of global stability. One may conjecture that for $a > 1$ the attractor of all slowly oscillating solutions of Eq. (1.1) coincides with the compact disk $\bar{W}$. A proof should involve a uniqueness result for slowly oscillating periodic solutions. A fact which is not very difficult to obtain is that necessarily any slowly oscillating periodic solution with $x(0) = 0$ has the symmetry

$$x(\cdot) = -x(\cdot + z)$$

with z the smallest zero in $(0, \infty)$. This might help to establish uniqueness. - Further one might expect that the disk $\bar{W}$ is a C^1-submanifold with boundary (as its analogue in [16, 17, 19]).

Another challenging problem is the analysis of the local bifurcation at $a = 1$: The particular unfolding of a piecewise smooth collage of two Bogdanov–Takens Singularities seems to exhibit the birth of a periodic orbit which is not of Hopf type, with period large and amplitude small for the parameter close to criticality. Work on this is in progress.

In a forthcoming paper [18] it is shown that for $a \to \infty$ the periodic solutions obtained in Theorem 7.1, after rescaling, converge to a square wave.

ACKNOWLEDGMENTS

Parts of the paper originate in the diploma thesis of A. E. from the program "Mathematics of Economics and Finance" at Comenius University in Bratislava. The work of P.B was supported partially by the VEGA Grants Nr. 1/0259/03 and 1/9155/02.

REFERENCES

1. Brauer, F., and Nohel, J. (1969). *The Qualitative Theory of Ordinary Differential Equations*, Benjamin, New York.
2. Brunovský, P., Erdélyi, A., and Walther, H. O. Short-term fluctuations of exchange rates driven by expectations. *In preparation.*
3. Cao, Y. (1996). Uniqueness of periodic solutions for differential delay equations. J. Diff. Eq. **128**, 46–57.
4. DeGrauwe, P., and Grimaldi, M. The exchange rate and its fundamentals. A choatic perspective. *CESifo Working Paper No. 639 (6),* January 2002.
5. Diekmann, O., van Gils, S., Verduyn Lunel S. M., and Walther, H. O. (1995). *Delay Equations: Functional-, Complex- , and Nonlinear Analysis,* Springer, New York.
6. Erdélyi, A. (2003). *A delay differential equation model of oscillations of exchange rates.* Diploma thesis, Bratislava.
7. Hale, J. K. (1977). *Theory of Functional Differential Equations*, Springer, New York.
8. Hale, J. K. (1985). Flows on centre manifolds for scalar functional differential equations. *Proc. R. Soc. Edinburgh* **101 A**, 193–201.
9. Hale, J. K., and Verduyn Lunel, S. M. (1993). *Introduction to Functional Differential Equations*, Springer, New York.
10. Kaplan, J. L., and Yorke, J. A. (1975). On the stability of a periodic solution of a differential delay equation. *SIAM J. Math. Anal.* **6**, 268–282.
11. Kaplan, J. L., and Yorke, J. A. (1977). On the nonlinear differential delay equation $x'(t) = -f(x(t), x(t-1))$. *J. Diff. Eq.* **23**, 293–314.
12. Krisztin, T., Walther, H. O., and Wu. J. (1999). *Shape, Smoothness and Invariant Stratification of an Attracting Set for Delayed Monotone Positive Feedback,* Fields Inst. Monographs, vol. 11, A.M.S., Providence.
13. Kuznetsov, Yu. A. (1995). *Elements of Applied Bifurcation Theory,* Springer, New York.

14. Mallet-Paret, J., and Sell, G. (1996). The Poincaré-Bendixson theorem for monotone cyclic feedback systems with delay. *J. Diff. Eq.* **125**, 441–489.
15. Nussbaum, R. D. (1979). Uniqueness and non-uniqueness for periodic solutions of $x'(t) = -g(x(t-1))$. *J. Diff. Eq.* **34**(2), 5–54.
16. Walther, H. O. (1991). An invariant manifold of slowly oscillating solutions for $\dot{x}(t) = -\mu x(t) + f(x(t-1))$. *J. reine angew. Math.* **414**, 67–112.
17. Walther, H. O. (1995). The 2-dimensional attractor of $x'(t) = -\mu x(t) + f(x(t-1))$. *Memoirs of the A.M.S.* **544**, 1–76.
18. Walther, H. O. (2004). Convergence to square waves for a price Model with delay. Submitted.
19. Walther, H. O., and Yebdri, M. (1997). Smoothness of the attractor of almost all solutions of a delay differential equation. DISS. MATH. **368**, 1–72.
20. Wright, E. M. (1995). A non-linear differential-difference equation. *J. reine angew. Math.* **194**, 66–87.

P. Brunovský

O minimách a maximách a o ich hľadaní I. O minimách a maximách

Matematické obzory 8 (1975), 43–50.

MATEMATICKÉ OBZORY 8/1975

O MINIMÁCH A MAXIMÁCH A O ICH HĽADANÍ

PAVOL BRUNOVSKÝ, Bratislava

I. O minimách a maximách

Človek sa často dostáva do situácie, v ktorej sa musí rozhodovať. Predmetom takéhoto rozhodovania často býva voľba nejakého naj. Môže tu ísť o drobné problémy (odtrhnúť si najčervenšie jablčko), závažnejšie (nájsť najkratšie spojenie s Kecerovských Peklian do Caracasu), až po veľmi vážne (výber najvhodnejšieho partnera pre život). Učene tomu hovoríme, že človek optimalizuje. Často sa mu pritrafí, že musí optimalizovať aj pri svojej práci. A tu musí riešiť úlohy niekedy jednoduché (z 57 časov dosiahnutých na Veľkej cene Slovenska vybrať najkratší), zložitejšie (vystrihnúť sako z najmenšieho kusa látky) a ťažké (určiť program letu rakety tak, aby vyniesla družicu na obežnú dráhu pri minimálnej spotrebe paliva).

Treba povedať, že človek je v tejto disciplíne majster. Myslím, že vo väčšine činností, ktoré človek dlhší čas robí, jeho postup je veľmi blízky optimálnemu. Ako príklad možno uviesť úlohu dopraviť električku z miesta A do miesta B po rovnej trati za daný čas pri minimálnej spotrebe energie. Dnešný matematický aparát umožnuje spočítať, že optimálny režim pripúšťa iba 4 fázy: maximálny ťah, udržovanie konštantnej rýchlosti, výbeh a maximálne brzdenie. A všimli ste si, ako postupuje vodič električky? Tak isto si myslím, že keby sa spočítalo, ako optimálne strihať látku na sako, sotva by sa výsledok veľmi líšil od toho, ako to krajčíri robia desaťročia. (Mimochodom, matematický popis tejto úlohy by bol ďaleko od jednoduchého!) Na tom nič nemení ani štatistika rozvodových súdov, ktorá hovorí, že, žiaľ, práve v jednom z najzávažnejších rozhodnutí, ktoré človek vo svojom živote musí urobiť, nie veľmi uspieva...

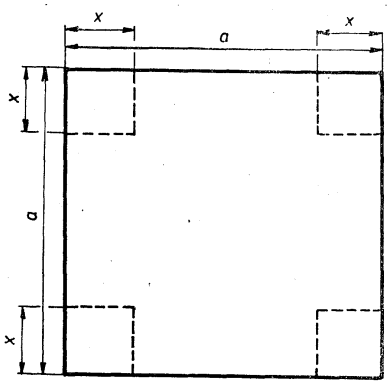
Napriek svojmu majstrovstvu sa človek často potrebuje obrátiť o pomoc na matematiku, napríklad preto, že nemá možnosť rozhodnutie opakovať dostatočne veľa ráz, aby sa k optimálnemu priblížil (raketa). Niekedy sa problém dá formulovať matematicky pomerne ľahko (raketa), inokedy ťažšie (strihanie látky) a niekedy to asi prakticky nejde (výber partnera pre život — našťastie, predstavte si tú bitku o toho najvhodnejšieho!).

Ak si zalistujeme trocha v spomienkach, možno si spomenieme, že jeden z prvých príkladov toho, že diferenciálny počet „na niečo je" sa týkal práve problému hľadania minima, resp. maxima reálnej funkcie reálnej premennej. Konkrétne, používali sme

Tvrdenie 1. Nech f: $(-\infty, \infty) \to (-\infty, \infty)$ je diferencovateľná. Aby funkcia f nadobúdala v bode $\hat{x}$ (lokálny) extrém (t. j. minimum alebo maximum), musí platiť $f'(\hat{x}) = 0$.

Školským príkladom pre použitie tejto vety je úloha odstrihnúť zo štvorca so stranou a rohy tak, aby sa zo zvyšku zložila škatuľa s maximálnym objemom (*obr. 1*). Ak si označíme x dĺžku strany odstrihnutého štvorca a $f(x)$ objem vzniknutej škatule, dostávame

$$f(x) = (a - 2x)^2 x = 4x^3 - 4ax^2 + a^2 x.$$



Obr. 1

Aby f nadobúdala v bode $\hat{x}$ maximum, musí platiť $f'(\hat{x}) = 0$, z čoho dostávame pre určenie $\hat{x}$ rovnicu

$$12\hat{x}^2 - 8a\hat{x} + a^2 = 0$$

a jej riešením dostávame dvoch „kandidátov" na lokálne extrémy

$$\hat{x}_1 = 1/6a, \qquad \hat{x}_2 = 1/2a.$$

Ľahko sa presvedčíme — napríklad tým, že spočítame $f(0)$ (prečo?) — že riešením úlohy je $\hat{x}_1$.

Takýchto pár príkladov sme si v knižke prečítali, prípadne na cvičení prerátali a ostal v nás dobrý pocit, že sme sa presvedčili o užitočnosti diferenciálneho počtu a že v prípade potreby hravo nájdeme extrém funkcie — aspoň jednej premennej. Ani sme si pritom nestačili uvedomiť, že tento a ďalšie príklady boli tak starostlivo vybraté, že sme mali z pekla šťastie v tom, že:

1. extrém sa nachádzal vnútri uvažovanej oblasti (v našom prípade $0 \leqq x \leqq a/2$),

2. lokálny extrém bol súčasne globálnym extrémom
a najmä, že

3. sme poznali analytický výraz pre f a rovnica $f'(\hat{x}) = 0$ sa dala explicitne riešiť.

Nebolo by treba dlho hľadať príklady, ktoré niektorú z týchto vlastností nemajú (napríklad triviálna úloha odrezať z danej palice palicu maximálnej dĺžky, pokiaľ pripustíme „prázdne" rezanie, nemá prvú vlastnosť). Pre úlohy, o ktorých nevieme vopred, že majú prvé dve vlastnosti, môžeme však ľahko na základe našich znalostí sformulovať

Tvrdenie 2. Nech f je diferencovateľná na intervale $[a, b]$ (rozumieme tým, že f má v bodoch a, b deriváciu sprava, resp. zľava a v každom vnútornom bode intervalu obojstrannú deriváciu) a f dosahuje v bode $\hat{x}$ lokálne minimum, potom buď $\hat{x} = a$ a $f'(a) \geqq 0$, alebo $\hat{x} = b$ a $f'(b) \leqq 0$, alebo $\hat{x} \in (a, b)$ a $f'(\hat{x}) = 0$[1]).

Tvrdenie 2 problém hľadania globálneho (t. j. nie iba lokálneho) minima, samozrejme, úplne nerieši, nemožno ho však považovať za zbytočné. Jeho zmysel spočíva okrem iného v tom, že silne obmedzuje množinu bodov, v ktorých f môže nadobúdať minimum. Dá sa dokonca celkom exaktne dokázať, že v istom zmysle pre „skoro všetky" funkcie je počet takýchto bodov konečný[2]).

Uvažovaním ohraničení stratila jednoduchá formulácia nutnej podmienky minima tvrdenia 1 na kráse. Matematik má oveľa radšej, ak môže alternatívnu formuláciu so slovami buď, alebo nahradiť jednotnou. Treba tu mať pritom na mysli, že formulácia analógie tvrdenia 1, resp. tvrdenia 2 pre funkcie viac premenných sa uvažovaním ohraničení stane ešte oveľa komplikovanejšou, pretože tu už zďaleka nevystačíme s intervalmi ako množinami, na ktorých sa minimalizuje.

Na pomoc si tu treba vziať jednoduchú myšlienku z praxe: Ak chceme, aby niekto nevkročil do zakázaného priestoru (v našom prípade mimo intervalu $[a, b]$), tak buď priestor oplotíme, alebo budeme za porušenie zákazu vyberať pokutu. Keďže nám veľmi záleží na tom, aby nikto zákaz neporušil, budeme vyberať pokutu poriadne vysokú; ako matematici máme v tomto smere výhodu v tom, že si môžeme dovoliť vyberať pokutu nekonečnú.

---

[1]) Obdobnú vetu možno, samozrejme, sformulovať pre maximum. Pretože však $\max f = -\min(-f)$, budeme v ďalšom hovoriť iba o minimách.

[2]) Presnejšie, dá sa dokázať, že pri vhodnej, prirodzeným spôsobom zvolenej topológii v priestore diferencovateľných funkcií na konečnom uzavretom intervale je množina funkcií, ktoré uvedenú vlastnosť nemajú, I. Baireovej kategórie.

Ináč povedané, interpretujeme f ako „cenu" a namiesto toho, aby sme skúmali funkciu f na intervale $[a, b]$, budeme skúmať funkciu $\tilde{f}: (-\infty, \infty) \to (-\infty, \infty]$, definovanú takto:

$$\tilde{f}(x) = \begin{cases} f(x) & \text{pre } x \in [a, b] \\ \infty & \text{pre } x \notin [a, b] \end{cases}$$

Pretože sme rozšírili priestor obrazov na $(-\infty, \infty]$ musíme povedať, ako v tomto priestore budeme počítať. Pre naše účely nám stačí definovať

$$x + \infty = \infty + x = \infty, \quad x \leqq \infty$$

pre každé $x \in (-\infty, \infty]$.

Je teraz zrejmé, že ak f je funkcia s reálnymi hodnotami na $[a, b]$, $f(\hat{x}) = \min_{x \in (-\infty, \infty)} f(x)$ vtedy a len vtedy, ak

$\hat{x} \in [a, b]$ a $f(\hat{x}) = \min_{x \in [a, b]} f(x)$ (prečo?).

S deriváciami, resp. diferenciálmi na takúto funkciu zrejme nemôžme (aj v prípade, že f je diferencovateľná na $[a, b]$, bude mať $\tilde{f}$ nespojitosti v bodoch a, b). Pomôže nám však pojem subdiferenciálu, ktorý sa definuje podobne ako diferenciál, s tým rozdielom, že rovnosť v definícii sa nahradí nerovnosťou.

Nech f: $(-\infty, \infty) \to (-\infty, \infty]$, $x \in (-\infty, \infty)$ a $f(x) < \infty$). Lineárnu funkciu $dy = a\,dx$ (reprezentovanú číslom a) budeme nazývať subdiferenciálom funkcie f v bode x, ak pre každú postupnosť $h_n \to 0$ existuje postupnosť $\{\omega_n\}$ taká, že $f(x + h_n) - f(x) \geqq ah_n + \omega_n$ pre každé n a $\lim_{n \to \infty} h_n^{-1}\omega_n = 0$.

Ekvivalentne môžeme subdiferenciál f v x definovať ako číslo a také, že platí

$$\liminf_{h \to 0} |h|^{-1}[f(x + h) - f(x) - ah] \geqq 0.$$

(dokážte!). Jedno-jednoznačné priradenie čísel a lineárnych funkcií nám umožňuje považovať subdiferenciál buď za číslo, alebo za lineárnu funkciu, podľa toho, ako nám to vyhovuje (pozri [1]).

Pre geometrickú predstavu nám poslúži ďalšia ekvivalentná definícia: a je subdiferenciál f v $\hat{x}$ vtedy a len vtedy, ak funkcia $F(x) = \min\{f(x), \hat{x} + a(x - \hat{x})\}$ je diferencovateľná v $\hat{x}$ a platí $F'(\hat{x}) = a$. (Nakreslite si graf funkcie F, ak $f(x) = x - x^2 \sin x$ a $\hat{x} = 0$!).

Je zrejmé, že subdiferenciál funkcie v niektorom bode nemusí existovať a tak isto, že takýchto subdiferenciálov môže byť viac. Označíme $\{f'\}(x)$ množinu subdiferenciálov f v x. Pre dobré pochopenie definície sa oplatí rozmyslieť si, že

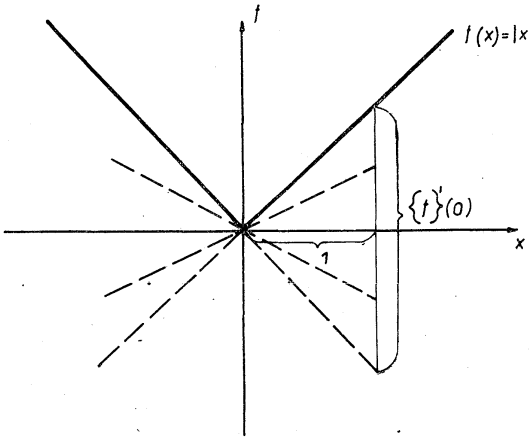
— pre $f(x) = |x|$ je $\{f'\}(x) = \{f'(x)\} = \operatorname{sign} x$ pre $x \neq 0$ a $\{f'\}(0) = [-1, +1]$ (*obr. 2*),

— pre $f(x) = -|x|$ je $\{f'\}(x) = -\operatorname{sign} x$ pre $x \neq 0$, ale $\{f'\}(0) = \emptyset$,

— $\{f'\}(x_0) = \{g'\}(x_0)$, ak $g(x) = f(x) + \alpha(x - x_0)^2$, alebo, všeobecnejšie, ak $\limsup_{h \to 0} |h|^{-1} |g(x_0 + h) - f(x_0 + h)| = 0$,

— ak $f(x) = \alpha$ pre $x \leqq 0$, $f(x) = \beta$ pre $x > 0$, kde $\alpha < \beta \leqq \infty$, potom $\{f'\}(0) = [0, \infty)$

a dokázať si to pomocou niektorej z ekvivalentných definícií.



Obr. 2

Podaktoré z vlastností, ktoré si pozorný čitateľ všimol na príkladoch, platia všeobecne a sú zhrnuté v tvrdení 3.

Tvrdenie 3. Nech f: $(-\infty, \infty) \to (-\infty, \infty]$. Potom

1. $\{f'\}(x)$ je uzavretý (prípadne prázdny) interval.

2. Ak f má deriváciu sprava a zľava v x (označíme ich $f'^{+}(x)$, resp. $f'^{-}(x)$), potom $\{f'\}(x) = [f'^{-}(x), f'^{+}(x)]$ (rozumieme $[a, b] = \emptyset$, ak $b < a$).

3. Ak $f(x) = \infty$ pre $x \in (x_0, x_0 + \varepsilon)$, $\varepsilon > 0$ a $f'^{-}(x)$ existuje, potom $\{f'\}(x_0) = [f'^{-}(x_0), \infty)$.

Ako dôsledok 2. bodu tvrdenia 3 dostávame, že $\{f'\}(x) = \{f'(x)\}$, ak f je diferencovateľná v bode x.

Aby sme dokázali 1. bod tvrdenia 3, treba nám ukázať, že $\{f'\}(x)$ je interval, t. j. že ak $\alpha, \beta \in \{f'\}(x)$ a $\gamma \in [\alpha, \beta]$, potom $\gamma \in \{f'\}(x)$ a že $\{f'\}(x)$ je zhora uzavretá množina, t. j. že ak $\alpha_n \to \alpha$, $\alpha_n \leqq \alpha$, $\alpha_n \in \{f'\}(x)$, potom aj $\alpha \in \{f'\}(x)$ (uzavretosť zdola sa totiž dokazuje úplne obdobne).

Nech teda α, $\beta \leqq \{f'\}(x)$. Nech $h_n \to 0$. Potom existujú postupnosti $\{\omega_n^i\}_{n=1}^{\infty}$ také, že $\lim h_n^{-1}\omega_n^i = 0$ a

$$f(x + h_n) - f(x) \geqq ih_n + \omega_n^i, \quad i = \alpha, \beta.$$

Položme $\omega_n = \omega_n^\beta$, ak $h_n \geqq 0$, $\omega_n = \omega_n^\alpha$, ak $h_n < 0$. Potom zrejme platí $h_n^{-1}\omega_n \to 0$ a ak $\alpha \leqq \gamma \leqq \beta$, potom

$$f(x + h_n) - f(x) \geq \beta h_n + \omega_n \geq \gamma h_n + \omega_n, \quad \text{ak} \quad h_n \geqq 0,$$
$$f(x + h_n) - f(x) \geq \alpha h_n + \omega_n \geq \gamma h_n + \omega_n, \quad \text{ak} \quad h_n < 0,$$

z čoho vyplýva $\gamma \in \{f'\}(x)$.

Pre dôkaz uzavretosti zhora použijeme druhú definíciu subdiferenciálu. Nech $\alpha_n \to \alpha$, $\alpha_n \leqq \alpha$, $\alpha_n \in \{f'\}(x)$ a predpokladajme $\alpha \notin \{f'\}(x)$. Potom platí

$$\eta = \liminf_{h \to 0} |h|^{-1}[f(x + h) - f(x) - \alpha h] < 0.$$

Z toho vyplýva, že existuje ľubovoľne malé $|h|$ také, že

$$|h|^{-1}[f(x + h) - f(x) - \alpha h] < 1/2\eta. \tag{1}$$

Pre dosť veľké n je $0 \leqq \alpha - \alpha_n \leqq 1/4\eta$, takže pre každé takéto $h > 0$ dostávame z (1) pre n dosť veľké

$$|h|^{-1}[f(x + h) - f(x) - \alpha h] < 1/4\eta + 1/4\eta \leqq 1/4\eta - h \,.\, h^{-1}(\alpha - \alpha_n)$$
$$|h|^{-1}[f(x + h) - f(x) - \alpha_n h] < 1/4\eta$$

(pre $h < 0$ vyplýva posledná nerovnosť z (1) triviálne).

Z toho dostávame pre dosť veľké n

$$\liminf_{h \to 0} |h|^{-1}[f(x + h) - f(x) - \alpha_n h] < 0$$

a to je v spore s predpokladom.

Vzhľadom na to, že už vieme, že $\{f'\}(x)$ je interval, na to, aby sme dokázali, že platí tvrdenie 3.2, stačí nám dokázať, že ak $f'^-(x) \leqq f'^+(x)$, potom $f'^+(x) \in \in \{f'\}(x)$ a že ak $\alpha > f'^+(x)$, potom $\alpha \notin \{f'\}(x)$ (dôkazy pre f'^- sú opäť obdobné). Z definície derivácie sprava, resp. zľava vyplýva

$$\lim_{h \to 0+} h^{-1}[f(x + h) - f(x) - f'^+(x)h] = 0, \tag{2}$$

$$\lim_{h \to 0-} |h|^{-1}[f(x + h) - f'^-(x) - f'^-(x)h] = 0. \tag{3}$$

Ak si uvedomíme, že pre $h < 0$ je $-f'^{-}(x)h < -f'^{+}(x)h$, dostaneme z (3)

$$\liminf_{h \to 0-} |h|^{-1}[f(x+h) - f(x) - f'^{+}(x)h] \geqq \lim_{h \to 0-} |h|^{-1}[f(x+h) - f(x) - f'^{-}(x)h] = 0.$$

Z toho a z (2) dostávame

$$\liminf_{h \to 0} |h|^{-1}[f(x+h) - f(x) - f'^{+}(x)h] \geqq 0,$$

a teda $f'^{+}(x) \in \{f'\}(x)$.

Ak $\alpha > f'^{+}(x)$, potom pre $h > 0$ dosť malé dostaneme z (2)

$$h^{-1}[f(x+h) - f(x) - f'^{+}(x)h] < 1/2(\alpha - f'^{+}(x)) = hh^{-1}(\alpha - f'^{+}(x)) - 1/2(\alpha - f'^{+}(x)),$$

$$h^{-1}[f(x+h) - f(x) - \alpha h] < -1/2(\alpha - f'^{+}(x)),$$

z čoho vyplýva

$$\liminf_{h \to 0} |h|^{-1}[f(x+h) - f(x) - \alpha h] < 0.$$

Keďže k dôkazu 3. bodu tvrdenia 3 netreba nijakú zvlášť novú myšlienku a dá sa vykonať tou istou technikou ako dôkazy prvých dvoch častí vety, prenecháva sa čitateľovi ako cvičenie.

Čo nám dáva pojem subdiferenciálu pre formuláciu nutnej podmienky minima, vidno z tvrdenia 4.

Tvrdenie 4. Ak $f\colon (-\infty, \infty) \to (-\infty, \infty]$ a $\hat{x}$ je lokálne minimum f, potom $0 \in \{f'\}(x)$.

Dôkaz je veľmi jednoduchý: ak $0 \in \{f'\}(\hat{x})$, potom

$$\liminf_{h \to 0} |h|^{-1}[f(\hat{x}+h) - f(\hat{x})] < 0,$$

čo je možné iba tak, že v každom okolí bodu 0 existuje h také, že $|h|^{-1}[f(\hat{x}+h) - f(\hat{x})] < 0$, t. j. $f(\hat{x}+h) < f(\hat{x})$, a teda $\hat{x}$ nie je lokálne minimum f.

Je zrejmé, že tvrdenie 4 obsahuje tvrdenie 2 v tom zmysle, že tvrdenie 2 je dôsledkom tvrdenia 4 v prípade, že funkcia f z tvrdenia 4 vznikne vlnovkovou operáciou z funkcie diferencovateľnej na danom intervale. Vyplýva to ihneď použitím 2. a 3. bodu tvrdenia 3 a dôkazy opäť prenechávame ako cvičenia čitateľovi.

Na záver tejto časti urobme si bilanciu toho, čo sme zavedením pojmu subdiferenciál získali. Z hľadiska „čisto" matematického sme získali nesporne dosť: podarilo sa nám sformulovať nutnú podmienku minima, ktorá

okrem toho, že je elegantnejšia než tvrdenie 2, je platná pre funkcie, na ktoré nie sú kladené nijaké podmienky regularity (to je v matematickej analýze prípad veľmi neobvyklý). Prakticista by zasa povedal, že sme sa dostali z dažďa pod odkvap, že nám totiž pojem subdiferenciálu pre samotné hľadanie minima nič nedáva: jednoducho možno subdiferenciál zistiť iba pre diferencovateľné funkcie a pre tie je tvrdenie 4 ekvivalentné tvrdeniu 2.

Nepochybne kus pravdy v tom je. Možno však úplne považovať eleganciu, všeobecnosť a jednoduchosť matematických tvrdení a teórií za samoúčelnú? Myslím si, že nie. Iste by bolo možné zaobísť sa bez znamienok $+$ a $-$ a opisovať tieto operácie slovne, ale myslím, že dnes sotva niekto pochybuje o tom, aký prevratný význam pre matematiku malo ich zavedenie. Samozrejme, nechcem tým povedať, že zavedenie pojmu subdiferenciál môže znamenať prínos, porovnateľný so zavedením $+$ a $-$; dokonca by som pripustil, že úžitok, ktorý z neho získame, nestojí za námahu — keby sme sa mohli obmedziť na funkcie jednej premennej. Vieme však, že s tým nevystačíme, a že matematika i ostatné vedy kladú pred nás optimalizačné úlohy, ktoré vedú na hľadanie extrémov funkcií viac premenných, dokonca funkcií na všeobecnejších priestoroch (raketa!). Tu už sa úžitok z pojmu subdiferenciál prejavuje oveľa vypuklejšie. Naviac má pojem subdiferenciálu hlboký význam, ktorý zatiaľ nemôžme vysvetliť, dostaneme sa však k nemu v tretej časti článku o kovexifikácii a dualite.

Pre funkcie n premenných $f: E^n \to (-\infty, \infty)$ analógiu tvrdenia 2 predstavujú tzv. Kuhnove — Tuckerove podmienky: Nech f, $g_i : E^n \to (-\infty, \infty)$, $i = 1, \ldots, m$ sú diferencovateľné a nech $\hat{x} \in K$ je taký bod, že $f(\hat{x}) = \min_{x \in K} f(\hat{x})$, $K = \{x \mid g_i(x) \leqq 0, i = 1, \ldots, m\}$. Označme $I(x) = \{i \mid g_i(x) = 0\}$. Predpokladajme, že vektory $\mathrm{d}g_i(\hat{x})$[3]), $i \in I(\hat{x})$ sú lineárne nezávislé (túto podmienku možno oslabiť, nie však úplne vypustiť). Potom existujú čísla $\mu_i \geqq 0$, $i \in I(\hat{x})$ (tzv. Kuhnove — Tuckerove multiplikátory) také, že

$$\mathrm{d}f(\hat{x}) = \sum_{i \in I(\hat{x})} \mu_i \, \mathrm{d}g_i \, (\hat{x}).$$

Ak prijmeme konvenciu $f(x) = \infty$ pre $x \notin K$, je táto pomerne zložito formulovaná veta obsiahnutá vo vete, ktorá znie rovnako ako tvrdenie 4, pričom subdiferenciál sa definuje doslovne rovnako ako pre funkciu jednej premennej.

To je všetko pekné, môžete povedať, uznávame, že subdiferenciál je užitočný, ale ako teda máme to minimum hľadať? O tom nabudúce.

Literatúra

[1] Brunovský P.: Derivácie a linearizácia. Matematické obzory 2/1972. Bratislava, ALFA 1972.

[3]) Rozumieme $\mathrm{d}g(x)$ ako vektor so zložkami $(\partial g/\partial x_j)\,(x)$, $j = 1, \ldots, n$.

P. Brunovský

O minimách a maximách a o ich hľadaní II. Ako sa množia králiky?

Matematické obzory 9 (1976), 29–38.

MATEMATICKÉ OBZORY 9/1976

O MINIMÁCH A MAXIMÁCH A O ICH HĽADANÍ

PAVOL BRUNOVSKÝ, Bratislava

II. Ako sa množia králiky

Tak teda, ako hľadať minimá? A čo to vôbec znamená?

Hádam netreba nikoho veľmi presvedčovať, že úlohy, v ktorých minimum možno presne vyrátať, budú skôr výnimkou než pravidlom. I keď sa podarí hodnotu, v ktorej sa minimum dosahuje, vyjadriť formulkou, nemusí to ešte znamenať, že skutočne možno minimum presne vypočítať — čo ak formulka napríklad obsahuje odmocniny?

Je teda hádam zrejmé, že nič nestratíme, ak nájdeme spôsob, ako minimum počítať iteratívne, s ľubovoľnou zadanou presnosťou[1].

Čo však znamená „zadaná presnosť"? Nájsť minimum funkcie f na intervale $[a, b]$ (ktorý v tejto časti budeme vždy predpokladať konečný, čo je v praxi vždy splnené) s presnosťou ε môžeme chápať vo viacerých zmysloch:

(x) nájsť interval $[x_\varepsilon^-, x_\varepsilon^+]$ taký, že $|x_\varepsilon^+ - x_\varepsilon^-| < \varepsilon$ a $\hat{x} \in [x_\varepsilon^-, x_\varepsilon^+]$ (kde, podobne ako v časti I, značí $\hat{x}$ bod, v ktorom f nadobúda minimum na $[a, b]$);

(xf) nájsť $x_\varepsilon \in [a, b]$ tak, že $|x_\varepsilon - \hat{x}| < \varepsilon$ a vypočítať $f(x_\varepsilon)$;

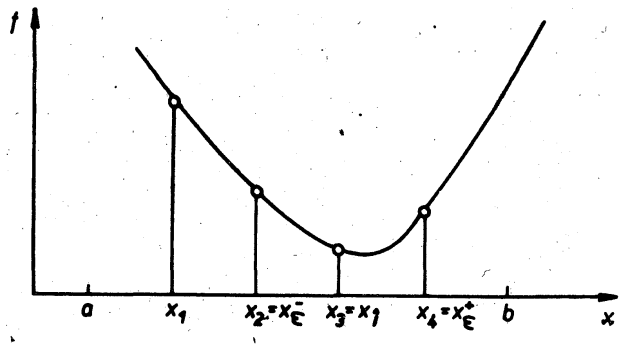
(f) nájsť x_ε tak, že $f(x_\varepsilon) - f(\hat{x}) < \varepsilon$ a vypočítať $f(x_\varepsilon)$.

Všetky z nich — a ešte možno ďalšie — majú zmysel a závisí od konkrétnej situácie, ktorá z nich je adekvátna.

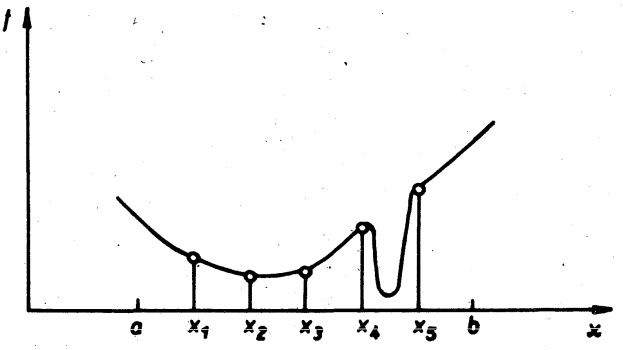
Pre riešenie všetkých týchto úloh máme ihneď naporúdzi jednoduchú „pasívnu" metódu:

Ak riešime úlohu (x), zvolíme k prirodzene tak, že $h = k^{-1}(b-a) < \varepsilon/2$, vypočítame hodnoty $f_i = f(x_i)$ v bodoch $x_i = a + ih$, $1 \leqq i \leqq k-1$ a položíme $x_\varepsilon^- = x_{\hat{i}-1}$, $x_\varepsilon^+ = x_{\hat{i}+1}$, kde $\hat{i}$ je také, že $f_{\hat{i}} = \min_{1 \leqq i \leqq k-1} f_i$; ak riešime úlohu (xf), zvolíme k tak, že $k^{-1}(b-a) < \varepsilon$ a položíme $x_\varepsilon = x_{\hat{i}}$. A konečne, ak riešime úlohu (f), volíme postupne $k = 2^p$, $p = 1, 2, \ldots$ nájdeme $f_{\hat{i}}$ pre $\hat{i} = \hat{i}(p)$ a pokračujeme až dovtedy, kým $f_{\hat{i}(p)} - f_{\hat{i}(r+1)}) < \varepsilon/2$ *(obr. 1).*

[1] Nech sa na mňa nehnevajú „computer scientisti" a štatistici, že tu abstrahujem od skutočnosti, že jednak zaokrúhľovacie chyby počítača a jednak nepresnosť východiskových dát dávajú na túto presnosť medze.



Obr. 1



Obr. 2

Pravda, hĺbavejší duch sa nad touto metódou zamyslí a bez väčšej námahy nájde príklad funkcie, pre ktorú metóda zlyhá, ako napr. na *obr. 2.* Skutočne, nemôžeme dokázať, že nám pasívna metóda vo všeobecnosti dáva riešenia úloh (x) až (f). Napriek tomu však „zdravý rozum" hovorí, že príroda nie je až taká zlomyseľná, že by pri dostatočne jemnom kroku bolo možné očakávať situáciu ako na *obr. 2* príliš často. V numerike sa napriek vyspelosti matematiky treba na zdravý rozum spoliehať dosť často — a napodiv úspešne ... A vôbec, komu sa nepáči, nech nájde lepšiu metódu, ktorá by bola tak všeobecná ako táto jednoduchá. Zaručujem mu, že ak bude úspešný, sláva ho neminie, lebo takej dosiaľ niet.

Nestálo by za to písať článok, ktorý by skončil takýmto tristným konštatovaním. Neboli by sme matematici, keby sme sa nesnažili dať zdravému rozumu nejakú exaktnú oporu, napríklad tým, že by sme dokázali oprávnenosť takejto jednoduchej metódy pre funkcie, splňujúce nejaké dodatočné predpoklady, ktoré, aj keď ich v praxi nie vždy môžme overiť, možno z rôznych príčin očakávať za splnené. Takými môžu byť napríklad ohraničenia na deriváciu f a pod. O to nám tu však nejde a ponechávame čitateľovi ako užitočné cvičenie zamyslieť sa nad tým. My sa

skôr zamyslíme nad tým, či nie je možné z výpočtov, ktoré musíme pri tejto jednoduchej metóde urobiť, voľačo ušetriť.

Ako sme už spomínali, vo všeobecnosti to nejde. Často sa však stáva, že funkciu f, ktorej minimum na $[a, b]$ hľadáme, môžeme na $[a, b]$ považovať za unimodálnu, pod čím myslíme, že f nadobúda na $[a, b]$ svoje minimum v jedinom bode, naľavo od tohto bodu je klesajúca a napravo od neho rastúca. Tak napr. funkcia z *obr. 1* je unimodálna, funkcia z *obr. 2* nie.

Hoci trieda unimodálnych funkcií je dosť úzka, možno často unimodalitu predpokladať z fyzikálnych dôvodov, alebo je možné jednoduchou metódou s hrubším krokom z pôvodne uvažovaného intervalu vymedziť interval, v ktorom je funkcia unimodálna.

Predpoklad unimodality umožňuje podstatne obmedziť počet bodov, v ktorých treba f počítať. Je totiž zrejmé, že ak $x_1 < x_2$ a $f(x_1) < f(x_2)$, potom f nemôže nadobúdať minimum napravo od x_2 (prečo?) a je teda zbytočné napríklad pokračovať v pasívnej metóde a počítať hodnoty v bodoch napravo od x_2. Táto úvaha naznačuje, že ak si nezvolíme vopred pevné body, v ktorých budeme f počítať, ale budeme každý ďalší bod voliť už s využitím informácie, získanej z predchádzajúcich výpočtov, môžeme voľačo získať.

A tu sa ukazuje matematický problém: K danému počtu n výpočtov funkcie f a danému intervalu $[a, b]$ nájsť optimálnu metódu postupnej voľby bodov ξ_k, $1 \leqq k \leqq n$ s využitím informácie o ξ_i a $f(\xi_i)$ $1 \leqq i \leqq k-1$. Ako vidno, optimalizovať možno nielen program letu rakety alebo strihanie škatule, ale aj samotný optimalizačný algoritmus.

Mohlo by sa zdať, že v dobe superrýchlych samočinných počítačov je smiešne baviť sa o tom, či vykonať 10 alebo 100 000 výpočtov funkcie f, veď čože je to pre taký počítač? Napriek tomu, dôvody tu sú. Tu, hľa, hneď tri:

1. čas potrebný na výpočet f môže byť skutočne dlhý,
2. nie vždy ide o počítanie — hodnoty f môžu byť výsledkom drahých meraní,
3. metódy hľadania extrémov funkcií viac premenných často obsahujú ako svoju časť hľadanie extrému funkcie jednej premennej, ktoré v procese výpočtu treba veľa ráz opakovať.

Na podopretie prvých dvoch dôvodov uvedieme dva príklady z „tvrdej" praxe. Za prvý vďačím svojej manželke, za druhý Ing. Brokešovi z Výskumného ústavu liehovarov a konzervární.

Niektoré pevné látky majú vlastnosť, že adsorbujú (pohlcujú) niektoré plyny. To sa využíva napríklad pri filtrácii exhalátov. Adsorpcia plynu v guľovej častici sa riadi tzv. II. Fickovým zákonom

$$\frac{\partial c}{\partial t} - \frac{1}{r^2}\frac{\partial}{\partial r}\left(a\frac{\partial c}{\partial r}\right) = 0$$

$$\frac{\partial c}{\partial r}(t, 0) = 0; \qquad c(t, r_0) = c_0; \qquad c(0, r) = 0$$

kde $c(t, r)$ značí koncentráciu adsorbátu (plynu) v mieste s polomerom r a v čase t a c_0 je koncentrácia na povrchu gule; podmienka $c(0, r) = 0$ značí, že na počiatku procesu je všade

v guli koncentrácii plynu nulová. Tzv. difúznu konštantu a však treba určiť experimentálne. Za tým účelom sa naplní kolóna guľovými časticami, do kolóny sa vháňa plyn, ktorý obsahuje ako svoju zložku adsorbát v koncentrácii c_0 a presným prístrojom — gravimatom — sa meria celkové naadsorbované množstvo do času t, $M(t)$, ktoré sa rovná množstvu naadsorbovanému v jednej častici $m(t)$ násobenému počtom častíc v kolóne q, t. j. $M(t) = qm(t) = q\int_0^r 4\pi\, cr^2 dr$. Konštanta a sa teraz hľadá tak, aby sa vypočítané naadsorbované množstvo $M_a(t)$ čo najlepšie zhodovalo s nameraným, pričom za kritérium zhody sa volí

$$f(a) = \int_0^T [M(t) - M_a(t)]^2 \,\mathrm{d}t = \int_0^T [M(t) - 4q\pi \int_0^{r_0} cr^2 dr]^2 \,\mathrm{d}t$$

Úloha teda vedie na hľadanie minima funkcie f na intervale $[0, \infty)$. K výpočtu každej hodnoty f však treba numericky vyriešiť parciálnu diferenciálnu rovnicu, čo nie je ani pre počítač celkom špás[2].

V druhom príklade ide o odlievanie membrán pre reverznú osmózu, jedným z ukazovateľov kvality ktorej je permeabilita (priepustnosť). Tá závisí od rýchlosti odlievania, avšak nik dnes nevie pre túto závislosť udať kvantitatívny vzťah. Čiže ostáva jediná možnosť, ak chceme permeabilitu maximalizovať: voliť rozličné odlievacie časy a merať výslednú permeabilitu.

Aby sme mohli úlohu optimalizácie optimalizačnej metódy presne formulovať, musíme sa rozhodnúť pre niektorú z úloh (x), (xf) a (f). Zvolíme si úlohu (xf), pretože pre ňu je riešenie najkrajšie. Skúste si rozmyslieť, ako je to s úlohami (x), (f). Výsledky vašich úvah si môžete skonfrontovať s [2].

Označme $\mathcal{F}(a, b)$ množinu unimodálnych funkcií na intervale $[a, b]$. Metódou n-krokovej postupnej minimalizácie Φ_n je daná postupnosťou funkcií $\xi_k^n = \xi_k^n(a, b, \xi_1, \ldots, \xi_{k-1}, \eta_1, \ldots, \eta_{k-1})$ s hodnotami v $[a, b]$, pomocou ktorých sa hodnoty ξ_k^n určujú rekurentným predpisom

$$\xi_k^n = \xi_k^n(a, b, \xi_1^n, \ldots, \xi_{k-1}^n, f(\xi_1^n), \ldots, f(\xi_{k-1}^n)),$$
$$k = 1, \ldots, n.$$

Úlohou je nájsť takú metódu $\hat{\Phi}_n$ (optimálnu), aby pre každý interval $[\alpha, b]$ a každú inú metódu Φ_n platilo

$$\max_{f \in \mathcal{F}(a,b)} \min_{1 \le i \le n} |\xi_i^n - \hat{x}| \le \max_{f \in \mathcal{F}(a,b)} \min_{1 \le i \le n} |\hat{\xi}_i^n - \hat{x}|$$

kde ξ_i^n, $\hat{\xi}_i^n$ sú po rade určené metódami Φ_n, $\hat{\Phi}_n$ (aj v ďalšom budeme bez ďalšieho dohovoru označovať body, vypočítané niektorou metódou rovnakými znakmi, ako samotnú metódu — napr. Φ_n^*, a_k^*).

Pre $n = 1$ je zrejmé, že optimálna metóda $\hat{\Phi}_1$ musí bodom a, b priraďovať stred intervalu $[a, b]$, t. j. $\hat{\xi}_1^1 = (a + b)$. Pri ľubovoľnej inej voľbe bodu ξ_1^1 je totiž vždy možné nájsť funkciu f takú, že $|\xi_1^1 - \hat{x}| > \frac{1}{2}(b - a)$ (ak $\xi_1^1 > \frac{1}{2}(a + b)$, je to $f(x) = x$, v opačnom prípade $f(x) = -x$).

[2] Úloha je tu trocha zjednodušená a na riešenie uvedenej rovnice je k dispozícii slušne konvergentný rad. V skutočnosti a závisí od c a vtedy sa rovnica stáva nelineárnou, čím sa jej riešenie skomplikuje.

Aj pre $n = 2$ je riešenie úlohy ľahké: znalosť funkcie v jednom bode nám ničím neprispieva pre lokalizovanie jej minima (dokážte!), preto je zrejmé, že voľba $\hat{\xi}_2^2$ bude nezávislá od hodnoty $f(\hat{\xi}_1^2)$. Obdobnou úvahou ako pre $n = 1$ zistíme, že musíme voliť $\hat{\xi}_1^2$, $\hat{\xi}_2^2$ tak, aby delili interval $[a, b]$ na tretiny (t. j. musí byť $\hat{x}_1 = a + \frac{1}{3}(b-a)$, $\hat{y}_1 = a + \frac{2}{3}(b-a)$, kde $\hat{x}_1 = \min\{\hat{\xi}_1^2, \hat{\xi}_2^2\}$, $\hat{y}_1 = \max\{\hat{\xi}_1^2, \hat{\xi}_2^2\}$).

Mýlil by sa však, kto by si myslel, že pre $n = 3$ je najvýhodnejšie deliť $[a, b]$ na štvrtiny; ak totiž poznáme hodnoty f v dvoch bodoch, vieme už zúžiť interval, na ktorom f môže minimum dosahovať.

Skúste teraz uhádnuť, ako riešiť úlohu pre $n = 3$! Ak ste hádali, že treba prvé dva body $\hat{\xi}_1^3$, $\hat{\xi}_2^3$ voliť tak, aby menší z nich bol bodom $\hat{y}_1$ metódy $\hat{\Phi}_2$ pre interval $[a, max\{\hat{\xi}_1^3, \hat{\xi}_2^3\}]$ a súčasne väčší z nich bodom $\hat{x}_1$ metódy $\hat{\Phi}_2$ pre interval $[\min\{\hat{\xi}_1^3, \hat{\xi}_2^3\}, b]$, hádali ste správne (nakreslite si to a určite, v akom pomere musia body $\hat{\xi}_1^3$, $\hat{\xi}_2^3$ deliť interval $[a, b]$, aby mali uvedené vlastnosti).

Dokazovať to zatiaľ nebudeme, pretože ukážeme, že toto pravidlo je časťou všeobecného pravidla, podľa ktorého sa bude riadiť voľba bodov $\hat{\xi}_k^n$ a ktoré pomenujeme pravidlo π.

Aby sme ho mohli formulovať pre všeobecné n, k, všimneme si najprv, že ak poznáme hodnoty $\xi_1^n, \ldots, \xi_k^n \in [a, b]$ a $f(\xi_1^n), f(\xi_2^n), \ldots, f(\xi_k^n)$, pre lokalizovanie minima funkcie f z toho môžeme vyťažiť nič viac a nič menej než to, že f nadobúda minimum na intervale $[a_k, b_k]$, kde a_k, b_k sú najbližší zľava, resp. sprava z bodov $a, b, \xi_1^n, \ldots, \xi_k^n$ k bodu ξ_{ik}^n takému, že $f(\xi_{ik}^n) = \min_{1 \le i \le k} f(\xi_i^n)$.[3] To znamená, že pokračovanie každého procesu postupnej n-krokovej minimalizácie Φ_n na $[a, b]$ spočíva v $n-k+1$-krokovej postupnej minimalizácii na $[a_k, b_k]$ so zadaným prvým bodom ξ_{ik}^n, ktorý označíme η_k^n: Z toho ihneď vyplýva, že η_{k+1} je jedným z bodov η_k, ξ_{k+1}^n. Z toho ďalej dostávame

$$[a_{k+1}, b_{k+1}] = \begin{cases} [a_k, y_k] & \text{ak } f(y_k) > f(x_k) \\ [x_k, b_k] & \text{ak } f(y_k) < f(x_k) \end{cases} \qquad (1)$$

kde $x_k = \min\{\eta_k, \xi_{k+1}^n\}$, $y_k = \max\{\eta_k, \xi_{k+1}^n\}$.

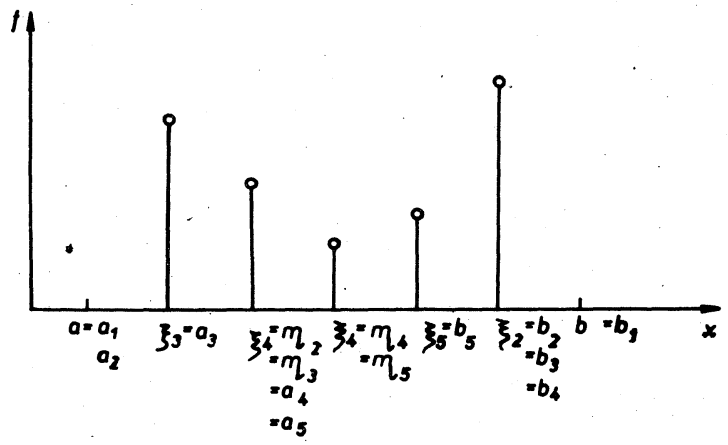
Pravidlo π teraz možno formulovať nasledovne:

Ak $[a_k, b_k]$ je interval, lokalizujúci minimum po k-tom kroku metódy Φ_n, potom body η_k, ξ_{k+1}^n sú totožné (nie nutne po rade) s prvými dvoma bodmi metódy Φ_{n-k+1} pre interval $[a_k, b_k]$.

Špeciálne prípady pre $n = 1, 2$, ako aj celková symetria úlohy vzhľadom na stred intervalu $[a, b]$ nás tiež oprávňujú očakávať, že pre ľubovoľné n a ľubovoľný interval $[a, b]$ sú $\hat{\xi}_1^n$ a $\hat{\xi}_2^n$ symetrické podľa bodu $\frac{1}{2}(a+b)$. Z (1), pravidla π a tejto symetrie môžeme už zrátať dĺžky intervalov $[\hat{a}_k, \hat{b}_k]$. Ak totiž označíme $\tau = \hat{\eta}_n - \hat{a}_n$ $\hat{x}_k = \min\{\hat{\eta}_k, \hat{\eta}_{k+1}\}$, $\hat{y}_k = \max\{\hat{\eta}_k, \hat{\eta}_{k+1}\}$, potom z pravidla π najprv vyplýva

$$\hat{b}_n - \hat{a}_n = 2\tau \qquad (2)$$

[3] Tu i v ďalšom nechávame bokom v praxi sa nevyskytujúci prípad, že by hodnota f vyšla rovnaká v rozličných bodoch a prenechávame čitateľovi rozmyslieť si túto možnosť.



Obr. 3

Z (1) a symetrie vyplýva

$$\hat{b}_{k+1} - \hat{a}_{k+1} = \hat{y}_k - \hat{a}_k = \hat{b}_k - \hat{x}_k \,,$$
$$\hat{x}_k - \hat{a}_k = \hat{b}_k - \hat{y}_k$$

z čoho ďalej dostávame

$$\hat{b}_{k-1} - \hat{a}_{k-1} = \hat{b}_{k-1} - \hat{x}_{k-1} + \hat{x}_{k-1} - \hat{a}_{k-1} =$$
$$= \hat{b}_k - \hat{a}_k + \hat{b}_{k+1} - \hat{a}_{k+1} \qquad (3)$$

(nakreslite si obrázok!), pretože $[\hat{a}_{k+1}, \hat{b}_{k+1}]$ musí byť jedným z intervalov $[\hat{a}_k, \hat{x}_k]$ $[\hat{y}_k, \hat{b}_k]$, alebo im symetrických podľa $\frac{1}{2}(\hat{a}_k + \hat{y}_k)$, resp. $\frac{1}{2}(\hat{x}_k + \hat{b}_k)$. Teda, ak označíme $F_0 = 1$, $F_1 = 1$ a

$$\hat{b}_k - \hat{a}_k = F_{n-k+2}\tau \qquad (4)$$

dostávame z (2) a (3)

$$F_{k+1} = F_k + F_{k-1} \quad \text{pre } k \geqq 1 \,, \qquad (5)$$

čo znamená, že dĺžky intervalov $[\hat{a}_k, \hat{b}_k]$ brané v opačnom poradí a merané dĺžkou polovice posledného z nich, $\tau = \hat{\eta}_n - \hat{a}_n$ tvoria starú známu Fibonacciho postupnosť, ktorá je obvykle jedným z prvých troch príkladov postupností, s ktorými sa človek na hodinách matematiky stretne a ktorou Leonardo z Pisy, zvaný Fibonacci, už v 13. storočí popisoval proces množenia králikov.

Pretože pre n-krokovú metódu $\hat{\Phi}_n$ je $b - a = \hat{b}_1 - \hat{a}_1 = F_{n+1}\tau$, dostávame z toho $|\hat{\xi}_n^n - \hat{x}| \leqq \tau = F_{n+1}^{-1}(b-a)$. Teda, ak máme zaručiť, aby sme na konci procesu poznali hodnotu f v bode, ktorý je od bodu $\hat{x}$ vzdialený o nie viac ako ε, musíme voliť n tak, aby $F_{n+1}^{-1}(b-a) < \varepsilon$, t. j. aby $F_{n+1} > \varepsilon^{-1}(b-a)$.

Ako bude vlastne vyzerať postupné hľadanie podľa Fibonacciho metódy $\hat{\Phi}_n$, vyvodíme teraz z (1) a (4), z ktorých vyplýva

$$b-a=\hat{b}_1-\hat{a}_1=F_{n+1}\tau\,,$$
$$\hat{y}_1-a=b-\hat{x}_1=\hat{b}_2-\hat{a}_2=F_n\tau$$

z čoho

$$(b-\hat{x}_1)/(b-a)=(\hat{y}_1-a)/(b-a)=F_n/F_{n+1}$$
$$(\hat{x}_1-a)/(b-a)=(b-\hat{y}_1)/(b-a)=1-(\hat{y}_1-a)/(b-a)=$$
$$=1-F_n/F_{n+1}=(F_{n+1}-F_n)/F_{n+1}=F_{n-1}/F_{n+1}$$

Teda body $\hat{x}_1$, $-\hat{y}_1$ treba voliť tak, aby úsečky $[a, \hat{x}_1]$, $[a, b]$ boli v pomere $F_{n-1}:F_{n+1}$ a bod $\hat{y}_1$ s ním symetricky; ak $f(\hat{x}_1)>f(\hat{y}_1)$, zvolíme $[\hat{a}_2, \hat{b}_2]=[a, \hat{y}_1]$, $\hat{\eta}_2=\hat{x}_1$, v opačnom prípade $[\hat{a}_2, \hat{b}_2]=[\hat{x}_1, b)]$, $\hat{\eta}_2=\hat{y}_1$. V každom prípade teda bude platiť buď $(\hat{\eta}_2-\hat{a}_2)/(\hat{b}_2-\hat{a}_2)=F_{n-1}/F_n$, alebo $(\hat{b}_2-\hat{\eta}_2)/(\hat{b}_2-\hat{a}_2)=F_{n-1}/F_n$. Vo všeobecnosti v k-tom kroku ($k\leqq n-1$) budeme mať interval $[[\hat{a}_k, \hat{b}_k]$ a bod $\hat{\eta}_k\in[\hat{a}_k, \hat{b}_k]$ taký, že buď $(\hat{\eta}_k-\hat{a}_k)/(\hat{b}_k-\hat{a}_k)=F_{n-k+1}/F_{n-k+2}$, alebo $(\hat{b}_k-\hat{\eta})/(\hat{b}_k-\hat{a}_k)=F_{n-k+1}/F_{n-k+2}$. Doplníme tento bod symetrickým na dvojicu $\hat{x}_k$, $\hat{y}_k$ a položíme $[\hat{a}_{k+1}, \hat{b}_{k+1}]=[\hat{a}_k, \hat{y}_k]$, $\hat{\eta}_{k+1}=\hat{x}_k$, ak $f(\hat{x}_k)<f(\hat{y}_k)$ a $[\hat{a}_{k+1}, \hat{b}_{k+1}]=[\hat{x}_k, \hat{b}_k]$, $\hat{\eta}_{k+1}=\hat{y}_k$ v opačnom prípade. Pre $k=n$ bude pre bod $\hat{\eta}_n$ platiť $(\hat{\eta}_n-\hat{a}_n)/(\hat{b}_n-\hat{a}_n)=F_1/F_2=1/2$, a teda $\hat{\eta}_n$ bude stredom úsečky $[\hat{a}_n, \hat{b}_n]$. Oveľa jednoduchšie a prehľadnejšie než týmto slovným popisom možno metódu $\hat{\Phi}_n$ popísať pomocou blokovej schémy, obvyklej v programovaní *(obr. 4)*.

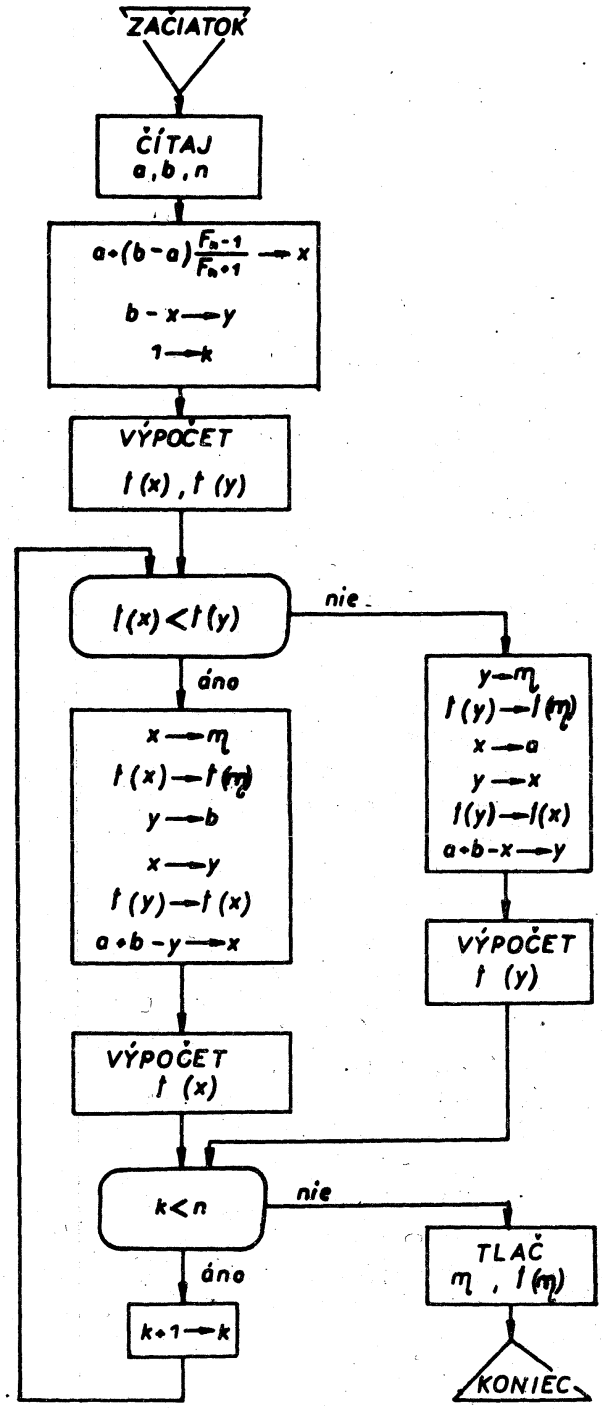
Týmto sme vlastne už ukázali, že je možné skonštruovať metódu, vyhovujúcu pravidlu π. Zostáva nám ešte dokázať, že metóda $\hat{\Phi}_n$ je skutočne optimálna n-kroková metóda postupného hľadania minima; súčasne dokážeme, že je jediná.

Dôkaz urobíme indukciou. Pre $n=1$ je tvrdenie zrejme pravdivé; predpokladajme, že je pravdivé aj pre $1\leqq k<n$.

Metóda $\hat{\Phi}_n$ dáva pre ľubovoľnú funkciu $|\hat{\eta}_n-\hat{x}|\leqq(b-a)/F_{n+1}$. Všimnime si, že rovnosť sa dosahuje okrem iného pre monotónne funkcie f — ak totiž napríklad f je rastúca, zrejme $\hat{a}_k=a$ pre všetky $1\leqq k\leqq n$, $\hat{x}=a$, a preto $\hat{\eta}_n-\hat{x}=\hat{\eta}_n-a=$ $=\hat{\eta}_n-\hat{a}_n=(b-a)/F_{n+1}$. Ukážeme, že k ľubovoľnej inej metóde Φ_n^* existuje funkcia, pre ktorú $|\hat{\eta}_n^*-\hat{x}|>(b-a)/F_{n+1}$, čím bude dokázané jednak že Φ_n je optimálna, jednak že je jediná.

Z indukčného predpokladu vyplýva, že nie je možné, aby sa Φ_n^* zhodovala s $\hat{\Phi}_n$ vo voľbe prvých dvoch bodov (ovplyvňujúcich a_2^*, b_2^*), ale líšila sa od $\hat{\Phi}_n$ v ďalších krokoch. Musí sa teda Φ_n^* líšiť od $\hat{\Phi}_n$ vo voľbe prvých dvoch bodov $x_1=\min\{\xi_1^n, \xi_2^n\}$, $y_1=\max\{\xi_1^n, \xi_2^n\}$. Predpokladajme $x_1^*\neq\hat{x}_1$ (prípad $y_1^*\neq\hat{y}_1$ je symetrický). Ak $x_1^*<\hat{x}_1$, potom pre klesajúcu funkciu f bude $[a_2^*, b_2^*]=[x_1^*, b]$, a teda $b_2^*-a_2^*<b-\hat{x}_1$; pretože podľa indukčného predpokladu Φ_n^* musí pokračovať ako Fibonacciho metóda $\hat{\Phi}_{n-1}$ na $[a_2^*, b_2^*]$, ktorá pre zvolenú funkciu f dáva $|\hat{\eta}_n-\hat{x}|=(b_2^*-a_2^*)/F_n>(b-\hat{x}_1)/F_n=(b-a)F_n/(F_nF_{n+1})=(b-a)/F_{n+1}\geqq$ $\geqq|\hat{\eta}_n-\hat{x}|/(b-a)$, Φ_n^* nemôže byť optimálna.

Ak $x_1^*>\hat{x}_1$, zvoľme f rastúcu na $[a, b]$. Potom bude zrejme $[a_2^*, b_2^*]=[a, y_1^*]$ a $[a_3^*, b_3^*]=[a, x_1^*]$, alebo $[a_3^*, b_3^*]=[a, \xi_3^{n*}]$ podľa toho, či zvolíme $\xi_3^{n*}<\hat{x}_1$ alebo


ZAČIATOK
ČÍTAJ
a, b, n
a+(b-a)Fn-1/Fn+1 → x
b - x → y
1 → k
VÝPOČET
f(x), f(y)
f(x)<f(y)
nie
áno
x → η
f(x) → f(η)
y → b
x → y
f(y) → f(x)
a+b-y → x
y → η
f(y) → f(η)
x → a
y → x
f(y) → f(x)
a+b-x → y
VÝPOČET
f(x)
VÝPOČET
f(y)
k<n
nie
áno
k+1 → k
TLAČ
η, f(η)
KONIEC

Obr. 4

$\xi_3^n{}^* > x_1^*$. V každom prípade však bude $b_3 - a_3 \geqq x_1^* - a > \hat{x}_1 - a = F_{n-1}/F_{n+1}(b-a)$. S pomocou indukčného predpokladu dostaneme opäť $|\eta_n^* - \hat{x}| > (b-a)/F_{n+1}$, čím je dôkaz ukončený.

Spočítajme si teraz, čo môžeme ušetriť pomocou metódy $\hat{\Phi}_n$ oproti pasívnej metóde pre unimodálnu funkciu. Pomocou formulky

$$F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right] \tag{6}$$

ktorá je známa pre Fibonacciho čísla (pozri [3]), si môžeme ľahko spočítať, že kým pre presnosť $\varepsilon = 10^{-3}\ (b-a)$ potrebujeme pri použití pasívnej metódy spočítať hodnoty f v 999 bodoch, pri použití metódy $\hat{\Phi}_n$ sa ich počet zredukuje na 13.

Formulka (6) má ešte aj iný význam. Pretože $1/2|1-\sqrt{5}| < 1$, možno z nej ľahko vyčítať, že platí

$$\lim_{n\to\infty} |F_n - z_n| = 0 \tag{7}$$

kde $z_n = (1/\sqrt{5})\ (1/2)\ (1+\sqrt{5})^n$. Jednoduchým výpočtom sa môžeme presvedčiť, že máme ešte aj to šťastie, že naviac má postupnosť $\{z_n\}$ spoločnú s $\{F_n\}$ aj vlastnosť (5), t. j. že platí $z_{n-1} + z_n = z_{n+1}$ pre $n > 1$. To ale znamená, že ju môžeme použiť na vytvorenie metódy postupného hľadania Φ^z podobne ako $\{F_n\}$, t. j. tak, že položíme $x_1^z = a + (z_{n-1}/z_{n+1})\ (b-a)$, $y_1^z = a + b - x_1^z = a + (1 - z_{n-1}/z_{n+1})\ (b-a) = a + (z_n/z_{n+1})\ (b-a)$. Proti $\{F_n\}$ má však postupnosť $\{z_n\}$ tú výhodu, že je geometrická, čo znamená, že $z_n/z_{n-1} = 1/2(1+\sqrt{5})$ nezávisí od n, a preto nezávisle od počtu krokov metódy a od toho, v ktorom kroku sme, delíme interval, lokalizujúci minimum v rovnakom pomere (preto ani v označení metódy Φ^z neindikujeme počet jej krokov). Asymptotická rovnosť (7) nám zaručuje, že pre veľké n metóda Φ^z nebude oveľa horšia ako $\hat{\Phi}_n$, pričom počítanie podľa nej je oveľa pohodlnejšie. Naviac, netreba dopredu vedieť počet krokov, ktoré chceme urobiť, a preto sa výhodne používa pre riešenie úlohy (f).

Metóde Φ^z sa tiež hovorí metóda zlatého rezu, pretože $z = 2(1+\sqrt{5})^{-1}$ je už zo staroveku známe ako pomer, v ktorom treba úsečku rozdeliť na dve časti tak, aby pomer dĺžky úsečky k dĺžke jej väčšej časti bol rovnaký ako pomer dĺžky väčšej časti k dĺžke menšej a ktorý bol významný v starovekej estetike.

Metódy $\hat{\Phi}_n$ a Φ^z nie sú zďaleka jediné, s ktorými možno postupne hľadať minimum unimodálnych funkcií. Napríklad, viaceré metódy sú vypracované pre prípad, že možno ľahko zisťovať deriváciu funkcie f. Nám tu však nešlo o prehľad metód minimalizácie, za ktorým čitateľ môže siahnuť do [2].

Tým sme konečne skončili s výpočtom toho, o čo nám tu nešlo. O čo nám tu teda šlo, alebo, prečo sme sa tu podrobne práve zaoberali Fibonacciho metódou? Pretože sa dá z nej vyvodiť niekoľko poučení. Po prvé, že aj v tých zdanlivo najprimitívnejších úlohách sa dá kadečo užitočného a zaujímavého vydumať. Po

druhé, že sa pritom často objavia celkom neočakávané súvislosti. A po tretie, metóda postupného hľadania je dobrým príkladom objektu, ktorého popis a analýza sa veľmi nešikovne formalizujú v obvyklom matematickom jazyku. V tomto smere môžeme ďakovať ľuďom od počítačov, že vytvorili symboliku a aparát (pozri napr. blokovú schému na *obr. 4*), ktoré sú pre takéto účely oveľa výhodnejšie.

Literatúra

[1] Brunovský, P.: O maximách a minimách a o ich hľadaní. In: Matematické obzory 8/75, Bratislava 1975, s. 43.
[2] Vasiljev, F. P.: Lekcii po metodam rešenija ektremaľnych zadač. Izd. Mosk. Univ. 1974.
[3] Vorobjev, N. N.: Čísla Fibonacci. Nauka 1969.

P. Brunovský

O minimách a maximách a o ich hľadaní. III. Zázrak a mystérium duality

Matematické obzory 10 (1976), 33–41.

MATEMATICKÉ OBZORY 10/1976

O MINIMÁCH A MAXIMÁCH A O ICH HĽADANÍ

PAVOL BRUNOVSKÝ

III. Zázrak a mystérium duality

Človek sa odjakživa snaží nachádzať v prírode harmóniu a poriadok a vytvárať ich vo svojich dielach. Máloktorá iná veda je taká bohatá na príklady ako matematika. Jedným z nich, ktorý vždy vo mne vzbudzuje údiv, je dualita vo všetkých jej rozličných podobách. O jednej z nich, súvisiacej s minimami a konvexitou, si voľačo povieme.

Ale začnime z iného konca — voľbou miss funkcie. Budeme hlasovať podľa osobných sympatií k typom funkcií. Na prvom mieste by sme sa vari zhodli — dúfam, že by ste naň tiež dali funkcie lineárne. Pokiaľ ide o druhé miesto, mnoho dôvodov — napr. výpočtové hľadisko — hovorí pre funkcie polynomiálne. Ja by som však hlasoval pre funkcie konvexné. Ak sa Vám to zdá teraz čudné, možno, že na konci článku budete mať pre mňa pochopenie.

Čo je konvexná funkcia, hádam viete: Funkciu f, definovanú na intervale I reálnej osi s reálnymi hodnotami, nazývame konvexnou, ak pre ľubovoľné $x, y \in I$ a ľubovoľné $\lambda \in [0, 1]$ platí $f(\lambda x + (1-\lambda)y) \leqq \lambda f(x) + (1-\lambda) f(y)$. Geometricky to znamená, že všetky body úseku grafu funkcie f medzi bodmi $(x, f(x))$ a $(y, f(y))$ ležia pod alebo na úsečke, spájajúcej tieto dva body (nakreslite si obrázok!).

Aby sme sa nezaťažovali technickými detailami, budeme hovoriť iba o konvexných funkciách, definovaných na celej priamke — ale pripustíme, aby nadobúdali hodnotu ∞ (s hodnotou ∞ budeme počítať ako v časti I). Množinu bodov, v ktorých f nadobúda konečné hodnoty, nazveme oblasťou konečnosti f a budeme predpokladať, že f je na svojej oblasti konečnosti spojitá.

Ako čiastočné vysvetlenie mojich sympatií ku konvexným funkciám môže slúžiť

Veta. 1. Nech f je konvexná. Potom

1. f nemá lokálne minimá okrem globálneho.
2. Oblasť konečnosti f je interval a v každom jeho vnútornom bode má f neprázdny subdiferenciál.
3. Pre každé $a \in \{f'\}(x)$ platí:

$$f(y) \geqq f(x) + a(y - x) \tag{1}$$

pre všetky y.

4. f nadobúda v bode $\hat{x}$ minimum vtedy a len vtedy, ak $0 \in \{f'\}(x)$.

Dôkaz 1. Predpokladajme, že v bode x funkcia f nadobúda lokálne minimum, ale nie globálne, teda že existuje taký bod y, v ktorom $f(y)<f(x)$. Pretože f je konvexné, platí pre $\lambda \in (0, 1)$

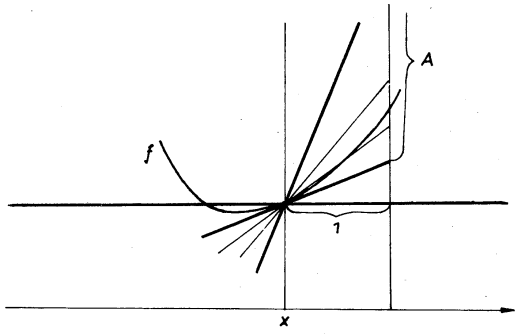
$$(2) \qquad f(x)+\lambda(y-x))=f(\lambda y+(1-\lambda)x)\leqq\lambda f(y)+$$
$$+(1-\lambda)\,.\,f(x)<\lambda f(x)+(1-\lambda)f(x)=f(x)\,.$$

Na druhej strane, pretože x je lokálne minimum, musí pre dosť malé $|\lambda|$ platiť $f(x)+\lambda(y-x))\geqq f(x)$, čo je v spore s (2). (Oveľa zrejmejšie vám to bude, ak si nakreslíte obrázok.)

2. Skutočnosť, že oblasť konečnosti je interval, dostaneme ihneď z toho, že ak $x\leqq y, f(x)<\infty, f(y)<\infty$, potom aj pre ľubovoľné $\lambda \in [0, 1]$ je $f(x+y\,(y-x))\leqq \lambda f(x)+(1-\lambda)f(y)<\infty$.

Dôkaz druhej časti je namáhavejší a ak sa vám do neho nechce, nakreslite si aspoň obrázok a dôkladne si premyslite, čo znamená. Uvidíte, že je to intuitívne zrejmá vlastnosť, ktorá spolu s (1) hovorí, že každým vnútorným bodom grafu funkcie f možno viesť opornú priamku — t. j. priamku, pod ktorou neleží nijaký bod grafu funkcie f. Táto hlboká vlastnosť konvexných funkcií sa bohato využíva v analýze, funkcionálnej analýze, nelineárnom programovaní a inde.

Nech teda x je vnútorný bod oblasti konečnosti f. Potom existujú také body y, z, že $y<x<z$ a $f(y)<\infty$, $f(z)\infty$. Označíme $a=\inf A$, kde $A=\{a\,|$ existuje $\xi>x$ také, že $f(\xi)<f(x)+\alpha(\xi-x)\}$. Inak povedané, a je infimum zo smerníc tých priamok p, prechádzajúcich bodom $(x, f(x))$, že vpravo od bodu x leží na grafe funkcie f bod pod p (pozri *obr. 1*).



Obr. 1.

Dokážeme, že a je konečné a $a \in \{f'\}(x)$. Je $f(z)<f(x)+f(z)-f(x)+1=$ $=f(x)+[(f(z)-f(x)+1)/(z-x)]\,(z-x)$, a teda $a\leqq(f(z)-f(x)+1)/(z-x)<$ $<\infty$. Predpokladajme $a=-\infty$. To znamená, že existuje postupnosť bodov $\{\xi_n\}$, $\xi_n>x$ taká, že $f(\xi_n)<f(x)-n(\xi_n-x)$. Označme $\lambda_n=(x-y)\,/\,(\xi_n-y)$.

Je $0<\lambda_n<1$ a z konvexity f vyplýva:

$$(3)\qquad f(x)=f(y+[(x-y)/(\xi_n-y)](\xi_n-y))=$$
$$=f(y+\lambda_n(\xi_n-y))=f(\lambda_n\xi_n+(1-\lambda_n)y)\leqq$$
$$\leqq\lambda_n f(\xi_n)+(1-\lambda_n)f(y)<\lambda_n f(x)+(1-\lambda_n)f(y)-$$
$$-\lambda_n n(\xi_n-x)=\lambda_n f(x)+(1-\lambda_n)f(y)-$$
$$-\lambda_n(\lambda_n^{-1}-1)\,n(x-y)=\lambda_n f(x)+(1-\lambda_n)[f(y)-n(x-y)]\,.$$

Pre dosť veľké n je $f(y)-n(x-y)<f(x)$, a preto z (3) vyplýva:

$$f(x)<\lambda_n f(x)+(1-\lambda_n)[f(y)-n(x-y)]<$$
$$<\lambda_n f(x)+(1-\lambda_n)f(x)=f(x)\,,$$

čo je nemožné.

Teda $a>-\infty$ a z jeho definície vyplýva, že existuje taká postupnosť bodov $\xi_n>x$, že $(f(\xi_n)-f(x))/(\xi_n-x)$ konverguje zhora, k a. Ďalej platí (1) pre všetky $y\geqq x$. Keby nie, existoval by totiž bod $y>x$ taký, že $f(y)-f(x)<a(y-x)$ a pre dosť malé $\delta>0$ by platilo $f(y)-f(x)<(a-\delta)(y-x)$, a teda $a-\delta\in A$, čo odporuje definícii a.

Ukážeme, že (1) platí i pre všetky $y\leqq x$. Predpokladajme opak, teda že existuje $\eta<x$ také, že $f(\eta)-f(x)<a(\eta-x)$. Pre dosť malé $\delta>0$ bude platiť

$$(4)\qquad f(\eta)-f(x)<(a+\delta)(\eta-x)$$

Pre dosť veľké n platí $(f(\xi_n)-f(x))/(\xi_n-x)<a+\delta$. Vyberme si takéto n a označme preň $\lambda=(x-\eta)/(\xi_n-\eta)$. Je $0<\lambda<1$ a podobne ako v (3) dostaneme:

$$f(x)<\lambda f(x)+(1-\lambda)[f(\eta)+(a+\delta)(x-\eta)]$$

Úpravou tejto nerovnosti dostaneme:

$$f(\eta)-f(x)>(a+\delta)(\eta-x)$$

čo odporuje (4).

3. Predpokladajme opak. Potom existuje taký bod y, že $f(y)<f(x)+(a-\delta)\times$ $\times(y-x)$. Potom však pre všetky $\lambda\in(0,1]$ platí $f(x+\lambda(y-x))\leqq\lambda f(y)+$ $+(1-\lambda)f(x)<\lambda[f(x)+(a-\delta)(y-x)]+(1-\lambda)f(x)=f(x)+\lambda(a-\delta)(y-x)$, čo znamená, že nemôže byť $a\in\{f'\}(x)$.

4. Toto tvrdenie je jedným smerom tvrdením 4 z časti I, druhým smerom zasa ihneď vyplýva z 1.

Keď už sú teda tie konvexné funkcie také sympatické, nebolo by od veci nájsť spôsob, ako z nekonvexnej funkcie urobiť konvexnú, ale tak, aby si pozmenená funkcia zachovala čo najviac z vlastností, ktoré sú pre nás dôležité. To sú, pravda,

predovšetkým minimá. Prirodzeným riešením tohto problému je najväčší konvexný dolný odhad, či „konvexná obálka" funkcie f, t. j. funkcia $\operatorname{co} f$ taká, že $\operatorname{co} f(x) \leqq f(x)$ pre všetky x a že ľubovoľná konvexná funkcia g je taká, že $g(x) \leqq f(x)$ pre všetky x (v ďalšom píšeme $g \leqq f$) spĺňa i $g \leqq \operatorname{co} f$ (porovnajte s definíciou uzáveru množiny ako najmenšej uzavretej množiny, obsahujúcej danú množinu, alebo supréma množiny ako jej najmenšieho horného ohraničenia).

Zatiaľ nemôžeme tvrdiť, že každá funkcia musí mať konvexnú obálku, jedno je však zrejmé, že konvexnou obálkou konvexnej funkcie je ona sama. Teraz si overíme, do akej miery konvexná obálka zachováva miminá, ako sme to od nej žiadali:

Veta 2. $\inf \operatorname{co} f = \inf f$; ak f nadobúda v bode $\hat{x}$ minimum, potom $f(\hat{x}) = \operatorname{co} f(\hat{x}) = \min \operatorname{co} f$.

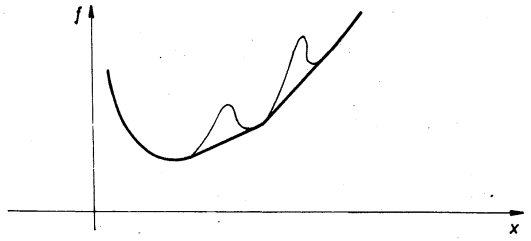
Vetu, prirodzene, treba chápať tak (pokiaľ nedokážeme existenciu obálky ku každej funkcii), že jej tvrdenie platí, ak obálka k danej funkcii existuje.

Dôkaz vety je založený na tom, že ak f, g sú konvexné, potom aj funkcia $\max \{f, g\}$ je konvexná (dokážte ako cvičenie). Teda aj $f = \max \{\inf f, \operatorname{co} f\}$ je konvexná funkcia; pretože platí $f \geqq \tilde{f} \geqq \operatorname{co} f$, musí platiť $\tilde{f} \leqq \operatorname{co} f$, z čoho vyplýva $\inf \operatorname{co} f = \inf f$. Ďalej platí $f(\hat{x}) = \operatorname{co} f(\hat{x})$, $\inf \operatorname{co} f = \inf f = f(\hat{x})$, a teda $\operatorname{co} f(\hat{x}) = f(\hat{x})$.

Príklad funkcie $f(x) = (1 + x^2)^{-1}$, ktorej $\operatorname{co} f \equiv 0$ ukazuje, že $\operatorname{co} f$ môže nadobúdať minimum aj vtedy, ak ho f nenadobúda. Ľahko sa možno presvedčiť, že tvrdenie vety možno preniesť aj na minimá na ohraničenom uzavretom intervale, kde už každá spojitá funkcia musí minimum dosahovať. Je zaujímavé, že kým túto vlastnosť si spojité funkcie nezachovávajú v nekonečnorozmerných priestoroch, konvexné spojité funkcie v mnohých prípadoch áno. Táto skutočnosť má veľký význam pre existenčné vety variačného počtu a teórie optimálneho riadenia a viedla k zavedeniu tzv. zovšeobecnených kriviek a relaxovaných riadení.

Ak si nakreslíte graf nejakej nekonvexnej funkcie, ruka vám takmer sama nakreslí jej konvexnú obálku (pozri *obr. 2*). Z obrázka možno ľahko vyčítať, ako funkciu $\operatorname{co} f$ určiť:

$$\operatorname{co} f(x) = \inf \{\lambda f(y) + (1 - \lambda) f(z) \mid \lambda y + (1 - \lambda) z = x, 0 \leqq \lambda \leqq 1\}$$



Obr. 2.

Ponechávame na čitateľovi, aby si vysvetlil geometrický význam tejto formulky a dokázal jej platnosť.

Iný spôsob určenia funkcie $\operatorname{co} f$ je analogický už spomínanému uzáveru množiny:

$$\operatorname{co} f(x) = \sup \{g(x) \mid g \text{ konvexná}, g \leqq f\}$$

My si však zvolíme inú cestu, ktorá sa sprvu bude zdať čudnou, napokon sa však ukáže prekvapujúco elegantnou a užitočnou.

Definujme najprv k ľubovoľnej funkcii f, ktorá má aspoň jednu konečnú hodnotu, duálnu funkciu f^* predpisom

$$f^*(\psi) = \sup_x H(x, \psi)$$

kde

$$H(x, \psi) = \psi x - f(x)$$

(písmeno H sa nezvolilo celkom náhodne, ale pre analógiu s Hamiltonovou funkciou vo variačnom počte, o ktorej bude reč neskôr).

K definíciám funkcií H a f treba ešte niečo dodať. Po prvé, keďže f môže nadobúdať aj hodnoty $+\infty$, môže sa v definícii funkcie H vyskytnúť $-\infty$, o čom sme nehovorili, ako s ním počítať. Pravidlá pre počítanie s ním sú však analogické pravidlám pre počítanie s $+\infty$. Ak by sme sa používaniu $-\infty$ chceli vyhnúť úplne, mohli by sme definovať $f^*(\psi) = -\inf_x [-H(x, \psi)]$. Po druhé, keďže pripúšťame iba funkcie s hodnotami v $(-\infty, \infty]$, musíme sa presvedčiť, že f^* nemôže nadobúdať hodnotu $-\infty$. Na tom však práve máme predpoklad o tom, že f nadobúda aspoň jednu konečnú hodnotu, o čom sa ľahko môžete presvedčiť.

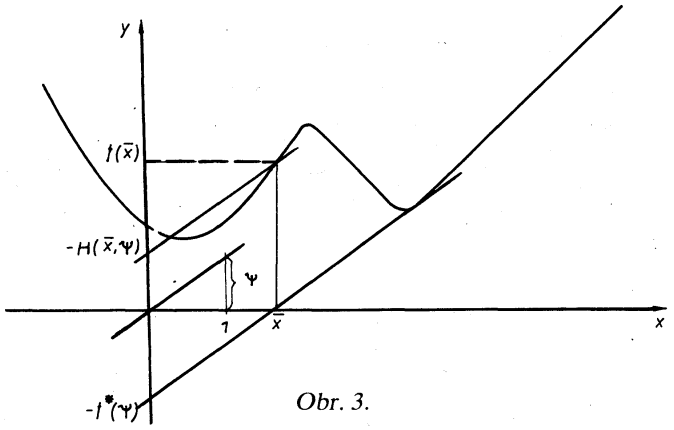
Ak chcete skutočne vniknúť do duality, bude dobré, ak si vypočítate f^* pre niekoľko funkcií, napr. pre už spomínanú funkciu $f(x) = (1+x)^{-1}$, ďalej pre funkcie $f(x) = ax + b$, $f(x) = |x|$, $f(x) = -|x|$, $f(x) = x^2$ a napokon $f(x) = 0$ pre $x = 0$ a ∞ pre $x \neq 0$.

Akú majú vlastne f^* a H geometrickú interpretáciu? Zvoľme ľubovoľný bod $(\bar{x}, f(\bar{x}))$ na grafe funkcie f a veďme ním priamku so smernicou ψ, ktorá bude mať rovnicu $y = \psi x + c$, kde $c = f(\bar{x}) - \psi\bar{x} = -H(\bar{x}, \psi)$ je jej úsek na osi y. Hodnota $f^*(\psi)$ je teda suprémum zo záporne vzatých úsekov na osi y, vyťatých priamkami, prechádzajúcich bodmi grafu funkcie f a majúcich smernicu ψ (*obr. 3*).

Dôležité vlastnosti duálnej funkcie zhrnieme do nasledujúcej vety:

Veta 3. Nech funkcia f sa nie všade rovná ∞. Potom f^* je konvexná funkcia. Ak existuje x_ψ také, že $f^*(\psi) = H(x_\psi, \psi)$, potom $\psi \in \{f'\}(x)$; ak f je konvexná, potom $\{f'\}(x) = \{\psi \mid x = x_\psi\} = \{\psi \mid f^*(\psi) = H(x, \psi)\}$.

Dôkaz konvexity f^* je veľmi jednoduchý. Pre ľubovoľné funkcie f, g a každé x platí totiž $f(x)+g(x)\leqq\sup f+\sup g$, a preto aj $\sup(f+g)\leqq\sup f+\sup g$. Vďaka tomu dostaneme pre $\lambda\in[0,1]$



Obr. 3.

$$f^*(\lambda\psi_1+(1-\lambda)\,\psi_2)=\sup_x\,[(\lambda\psi_1+(1-\lambda)\,\psi_2)\,x-f(x)]=$$

$$=\sup_x\,[(\lambda\psi_1+(1-\lambda)\,\psi_2)\,x-\lambda f(x)-(1-\lambda)\,f(x)]\leqq$$

$$\leqq\sup_x[\lambda\psi_1x-\lambda f(x)]+\sup_x[(1-\lambda)\,\psi_2x-(1-\lambda)\,f(x)]=$$

$$=\lambda\sup_x\,[\psi_1x-f(x)]+(1-\lambda)\sup_x\,[\psi_2x-f(x)]=$$

$$=\lambda f^*(\psi_1)+(1-\lambda)\,f^*(\psi_2)\,,$$

čo znamená, že f^* je konvexná.

Podľa definície x_ψ platí $H(x_\psi,\psi)\geqq H(x,\psi)$ pre všetky x, čo znamená $\psi x-f(x)\leqq\psi x_\psi-f(x_\psi)$, alebo $f(x)\geqq f(x_\psi)+\psi(x-x_\psi)$, a teda $\psi\in\{f'\}(x_\psi)$.

Aby sme dokončili dôkaz vety, stačí nám už iba dokázať, že z $\psi\in\{f'\}(x)$ vyplýva $x=x_\psi$. Ale to je vlastne tvrdenie 4 vety 1.

A teraz príde zlatý klinec programu: ak funkcia f^* sa nie všade rovná ∞, môžeme s ňou urobiť presne to isté, ako s f, t. j. $(f^*)^*$ (namiesto toho budeme písať f^{**}). Hádajte, čo dostaneme! Odpoveď dáva:

Veta 4. Nech $f<\infty$. Potom co f existuje práve vtedy, ak sa f^* nie všade rovná ∞; platí $f^{**}=\mathrm{co}\,f$.

Najkrajšie sa prejaví táto veta práve pri aplikácii na konvexné funkcie: ak $f<\infty$

je konvexná, potom $f^{**} = f$. K tomu si stačí uvedomiť, že tvrdenie 2 vety 1 spolu s vetou 3 zaručujú, že ak $f < \infty$ je konvexná, potom $f^* \not\equiv \infty$ (presvedčte sa!).

K dôkazu vety už máme kadečo pripravené, takže nebude ani taký zlý, ako by človek od takej prekvapujúcej vety mohol čakať. Pre každé ψ platí:

$$H_{f^*}(\psi, x) = x\psi - f^*(\psi) = x\psi - \sup_y H_f(y, \psi) =$$

$$= x\psi - \sup_y \{ \psi y - f(y) \} \leqq x\psi - [\psi x - f(x)] = f(x)\,.$$

z čoho

$$f^{**}(x) = \sup_\psi H_{f^*}(\psi, x) \leqq f(x)$$

(index pri H označuje, ku ktorej funkcii H patrí).

Nech teraz $\bar{f}$ je ľubovoľná konvexná funkcia taká, že $\bar{f} \leqq f$. Dokážeme, že potom platí aj $\bar{f} \leqq f^{**}$. Pretože z vety 3 vieme, že f^{**} je konvexná, bude tým dokázané

$$f^{**} = \operatorname{co} f$$

Zvoľme $\psi \in \{\bar{f}'\}(x)$ (pretože $\bar{f}$ je konvexná, existuje podľa vety 1. Potom podľa vety 3 platí:

$$\psi x - \bar{f}(x) = \sup_y \{ \psi y - \bar{f}(y) \} \geqq \sup_y \{ \psi y - f(y) \} = f^*(\psi) =$$

$$= \psi x - x\psi + f^*(\psi) \geqq \psi x - \sup_x \{x\chi - f^*(\chi)\} = \psi x - f^{**}(x)\,,$$

z čoho vyplýva $\bar{f}(x) \leqq f^{**}(x)$.

Na ukončenie dôkazu vety nám ešte treba dokázať, že ak existuje co f, potom f^* sa nie všade rovná ∞.

Pretože $\operatorname{co} f \leqq f < \infty$, má co f podľa vety 1 v každom bode x neprázdny subdiferenciál; zvoľme teda x a $\psi \in \{\operatorname{co} f'\}(x)$.
Podľa vety 1 to znamená:

$$\operatorname{co} f(y) \geqq \operatorname{co} f(x) + \psi(y - x)$$

pre všetky y, a teda aj

$$f^*(\psi) = \sup_y \{ \psi y - f(y) \} \leqq \sup_y \{ \psi y - \operatorname{co} f(y) \} \leqq \psi x - \operatorname{co} f(x) < \infty$$

Mohli by ste sa ešte právom opýtať, prečo sme zavádzali funkcie s hodnotami ∞,

keď sme hlavný výsledok nakoniec dokázali iba pre konečné funkcie. Nuž, je to preto, že aj pre konečnú funkciu môže vyjsť jej duálna s nekonečnými hodnotami (čo už viete, ak ste si vypočítali príklady). Okrem iného, veta platí aj za predpokladu, že f nadobúda aspoň jednu konečnú hodnotu, ak navyše predpokladáme, že je polospojitá zdola (šarapatu robia okraje intervalu konečnosti f^{**}, v ktorých môže byť $f^{**} < \mathrm{co}\, f$).

Ak sa vám ešte stále nechce veriť vete 4, preskúšajte si ju na konkrétnych príkladoch, ktoré sú uvedené pred vetou 3, alebo ktoré si sami vymyslíte. Ale aj keď tomu veríte, je to celkom zábavné. Že veta neplatí celkom všeobecne, môžete sa presvedčiť na príklade konvexnej funkcie $f(x) = 0$ pre $x \in (0, 1)$, 1 pre $x = \pm 1$ a ∞ pre $|x| > 1$.

Všimnime si ešte jednu vec, ktorá nám bude v ďalšom užitočná: Ak f je konvexná a konečná (a teda $f^{**} = f$), potom $f^{***} = (f^*)^{**} = (f^{**})^* = f^*$ už bez akýchkoľvek ďalších predpokladov na f^* (teda aj pre f^* nadobúdajúcu nekonečné hodnoty!).

Ako podivuhodne sa prepletajú vlastnosti funkcie a jej duálnej funkcie, ukazuje:

Veta 5. Nech f je konečná. Potom platí:

1. Ak f je konvexná, $\psi \in \{f'\}(x)$ vtedy a len vtedy, ak $x \in \{f^{*\prime}\}(\psi)$.

2. Ak f nadobúda v bode $\hat{x}$ minimum, potom $\hat{x} \in \{f^{*\prime}\}(0)$; ak f je konvexná, platí to aj opačným smerom.

Prv než vetu 5 dokážeme, uvedomme si geometrický význam jej tvrdenia 2: $f(\hat{x})$ je maximálny z úsekov, ktoré na priamke $x = \hat{x}$ vytnú oporné priamky grafu funkcie f.

Dôkaz. $x \in \{f^{*\prime}\}(\psi)$ znamená podľa vety 1 $f^*(\chi) \geqq f^*(\psi) + x(\chi - \psi)$, a teda $\psi x - f^*(\psi) \geqq \chi x - f^*(\chi)$ pre všetky χ. Preože podľa vety 4 $f^{**} = f$, znamená to

$$\psi x - f^*(\psi) = \max_{\chi} \{\chi x - f^*(\chi)\} = f^{**}(x) = f(x) =$$

$$\psi x - \sup_{y} \{\psi y - f(y)\} \geqq f(x)$$

$$\psi x \geqq f(x) + \sup_{y} \{\psi y - f(y)\} \geqq f(x) + \psi y - f(y)$$

pre každé y, a teda

$$f(y) \geqq f(x) + \psi(y - x)$$

čo znamená $\psi \in \{f'\}(x)$.

Všimnime si, že konečnosť funkcie f sme potrebovali iba k tomu, aby sme si zabezpečili splnenie rovnosti $f^{**} = f$ a že tvrdenie platí i za tohto predpokladu.

Pretože $f = f^{**}$, znamená $\psi \in \{f'\}(x)$ aj $\psi \in \{f^{**'}\}(x)$. Ak v tvrdení, ktoré sme už dokázali, zameníme f za f^* (čo môžeme vzhľadom na to, že $f^{***} = f^*$), dostaneme $x \in \{f^{*'}\}(\psi)$.

Ak f nadobúda v bode $\hat{x}$ minimum, potom podľa vety 2 nadobúda v bode x minimum i funkcia $\mathrm{co}\, f = f^{**}$ a platí $f^{**}(\hat{x}) = f(\hat{x})$. Podľa vety 1 platí $0 \in \{f^{**'}\}(\hat{x})$ a podľa 1. tvrdenia našej vety z toho vyplýva $\hat{x} \in \{f^{*'}\}(0)$ (opäť sme použili $f^{***} = f^*$!). Naopak, ak $\hat{x} \in \{f^{*'}\}(0)$, potom podľa 1. tvrdenia $0 \in \{f'\}(\hat{x})$ a podľa vety 1, ak f je konvexná, nadobúda v bode $\hat{x}$ minimum.

Veta 3 je tým zázrakom i mystériom. Zázrakom preto, že to tak krásne klape, a mystériom je, prečo to tak klape. Veta 3 je totiž jednou z mnohých analogických viet v geometrii, algebre, topológii i funkcionálnej analýze, pre ktoré niet spoločného vysvetlenia. Tým myslím, že ani nie sú „odpozorované z prírody", ani niet všeobecnej matematickej teórie, z ktorej by sa dali vyvodiť ako špeciálne prípady.

Pravda, opäť sa môžeme opýtať ústami prakticistu: Dobre, dualita je krásna, ale načo ju potrebujeme?

V tejto forme asi nanič. Ale kto sa stretol s viazanými extrémami v analýze, alebo s kanonickými premennými a Legendrovou transformáciou vo variačnom počte, alebo sa s nimi stretne, potvrdí mi, že sú to nie ľahké veci na pochopenie. A pritom si to pomocou funkcie H a premennej ψ môže v čistej geometrickej forme „nakresliť". No význam kanonických premenných v mechanike, či ich mladších súrodencov — duálnych premenných v lineárnom, čo v nelineárnom programovaní sotva niekto odškriepi.

Zaujímavé je, že vývoj tu nešiel od nášho jednoduchého geometrického modelu duality ku kanonickým premenným, ale práve naopak. Ťažko je vžiť sa dnes do toho, ako rozmýšľal Hamilton pred poldruha storočím, ale skôr sa zdá, že na svoje objavy prišiel formálnym počítaním, než z geometrických predstáv. Tým skôr mu treba vzdať úctu.

Pokiaľ ste z predchádzajúcich častí získali dojem, že nadŕžam matematike, ktorá priamo počíta, tak touto časťou som chcel vyjadriť hold matematike, ktorá je krásna, a matematike, ktorá vysvetľuje. História ukazuje, že zdôrazňovanie len jedného z aspektov vedie iba k degenerácii matematiky a že iba jej prirodzený komplexný rozvoj vedie k skutočnému pokroku.

P. Brunovský

Vektory očami nečistého matematika

Matematické obzory 16 (1980), 3–6.

MATEMATICKÉ OBZORY 16/1980

VEKTORY OČAMI NEČISTÉHO MATEMATIKA

alebo nezabudnime, že vektor je šípka, a to nie hocijaká. Ale taká, čo je všade a nikde. Asi tak, ako maličkí panáci na Chagallových obrazoch.

Pravdaže, vektor je n-tica čísel, trieda ekvivalencie, posunutie a neviem čo všetko iné, ako sme sa o tom dočítali na stránkach *Matematických obzorov* (1—5). Teda presnejšie, možno ho tak definovať a možno si ho aj tak predstavovať. Naproti tomu intuitívna *šípková* definícia vlastne vôbec nie je definíciou podľa našich súčasných kritérií. Ale zato je to veľmi užitočná a názorná predstava — trúfal by som si povedať, že v tomto smere má veľké prednosti pred všetkými presnými definíciami a že všetky spomínané definície sú tu vlastne preto, aby túto predstavu vyjadrili.

V čom vlastne spočíva užitočnosť a príťažlivosť šípkovej predstavy? Predovšetkým v tom, že je veľmi jednoduchá a názorná a pritom dostatočne výstižná na to, aby sme z nej mohli odvodiť veľmi veľa z toho, čo o vektoroch vieme. A čo je dôležité, hoci to nie je presná definícia, nikdy neodvodíme nič chybného.

Práve preto by sme nikdy nemali na túto predstavu zabudnúť a tobôž nie sa za ňu hanbiť.

Nejde tu však natoľko o vektor, ako o predstavy v matematike všeobecne.

Možno sa opýtať — načo o tom písať? Kto komu bráni, aby si čokoľvek akokoľvek predstavoval? Isteže, nikto nikomu nebráni. Ale ak chceme niekoho naučiť dobre maľovať, nestačí nebrániť mu v tom — treba ho voľačo aj naučiť. A tak aj v matematike treba žiakom a študentom vštepovať správne predstavy. Predstavy názorné a jednoduché, ktoré umožňujú voľačo vypočítať a voľačo vytvoriť. Neviem, či to nie je tak, že pri všetkých našich reformách, učebných plánoch atď. priveľmi nedbáme na to, aby sme sa, preboha nikde nedopustili nejakej logickej nedôslednosti, pričom nám utečie sem-tam názornosť. Ako keby sme sa báli, že ak si dovolíme čo len trocha oprieť o predstavu, postavia nás na pranier

a budú na nás prstom ukazovať ako na trestuhodných diletantov, alebo horšie, zradcov matematiky. Najmarkantnejšie sa to hádam prejavuje pri vyučovaní nematematikov na vysokých školách.

Teda, nezabudnime, že vektor je šípka. A patrilo by sa dodať — a rýchlosť. A dva razy to podčiarknuť. To preto, že ak nám asi trocha unikajú geometrické predstavy, o fyzikálnych je to celkom isté. Neviem presne, ako je to na strednej škole, ale je neodpustiteľným hriechom, že fyzika prakticky vymizla z učebných programov matematikov na vysokých školách. Nemali by sme zabudnúť, že veľká časť našej súčasnej matematiky je veľmi úzko spätá s fyzikou (diferenciálny a integrálny počet, lineárna algebra, variačný počet), s ktorou a pre ktorú sa vyvíjala od vynájdenia infinitezimálneho počtu pred viac ako 300 rokmi. Na čom sa má matematik učiť robiť matematické modely, ak nie na fyzikálnych úlohách?

Ale vráťme sa k vektoru. Na čom inom si môže človek tak krásne predstaviť vektor a potrebu jeho zavedenia, ako práve na vektore rýchlosti? Tam to úplne kričí, že kým *polohový vektor* vlastne vektorom nie je (ten musí byť niekde *upevnený* a sčitovať alebo prenášať takéto vektory nemá prirodzený zmysel), vektory rýchlostí dvoch hmotných bodov možno sčitovať, nech sú body akokoľvek ďaleko od seba. Kde by človek prišiel na taký čudný pojem, ako je vektorový súčin, nebyť, povedzme, elektromagnetického poľa vodiča a *pravidla pravej ruky?*

Napokon, ešte jedna vec stojí za zmienku. Predstava má veľmi dôležitú úlohu nielen v štádiu tvorby, ale aj pri dokazovaní a overovaní výsledkov — väčšiu, než sa to zvyčajne pripúšťa. Je známe, že starí páni infinitezimálneho počtu — Newton, Euler, d'Alembert a iní — sa nemýlili viac ako my, hoci nemali k dispozícii tak precízne deduktívne vybudovanú matematickú teóriu, akú máme dnes my. To sa často zdôvodňuje tým, že to boli géniovia. O tom iste nik nepochybuje, ale bude tu aj iný dôvod: mali veľmi bohaté a dobré predstavy — predovšetkým geometrické a fyzikálne. A to je veľmi mocný prostriedok na overovanie správnosti výsledkov — možno mocnejší, ako logicko-deduktívny. A vôbec, prečo ísť tak ďaleko do histórie: Je obdivuhodné, ako dobrí inžinieri alebo fyzici vedia vymýšľať nové matematické pojmy, vety a overovať si ich, mysliac pritom nie na prvé, či druhé derivácie, ale na rýchlosti, zrýchlenia, momenty a podobne. Alebo iný príklad: Definovať matematicky precízne tenzor nie je také ľahké a vyžaduje to pomerne vysokú matematickú erudíciu. Vie to

dosť málo vyštudovaných matematikov a ešte menej inžinierov alebo iných matematikov. Napriek tomu vedia inžinieri a mechanici virtuózne počítať s tenzormi — práve preto, že si za nimi predstavujú napätia, deformácie atď. A čo je prekvapujúce — vedia to neporovnateľne zručnejšie, ako matematici, ktorí síce tenzor definovať vedia, ale zodpovedajúce predstavy nemajú.

Keď hovoríme o predstavách, zaujímavý je aj prípad geniálneho indického matematika RAMANUJANA, ktorý vymýšľal fantasticky komplikované číselne-teoretické a kombinatorické identity, ktoré nevedel dokazovať, ani odvodiť. Domnievam sa, že o všetkých sa po kratšom alebo dlhšom čase dokázalo, že sú správne. Ich dokazovaním sa nezaoberal nik iný, ako HARDY — vari najväčšia postava teórie čísel 20. storočia. Ako mohol Ramanujan tie identity vymýšľať? Hádam mal nejaké svojské predstavy, ktoré my nemáme...

Zaujímavý je aj príklad teórie grafov. Je to disciplína, ktorá by sa úplne dala vybudovať bez grafov (vrcholy sú prvky množiny a dvojice týchto prvkov s istými vlastnosťami...). Veď keď sa grafické úlohy riešia na počítači, grafická predstava aj tak musí ísť bokom a takto človek vlastne grafy do počítača zapisuje. Netreba však azda nijako osobitne vysvetľovať, že teória grafov by bez svojej grafickej interpretácie sotva mohla existovať.

Napokon, či sa nám to páči, alebo nie, skutočnosť je taká, že aj konečnú kontrolu prakticky každého zložitejšieho matematického postupu si robíme nie logicko-deduktívne, ale akosi globálne. Čo to značí? Je to akési vnorenie výsledku do subjektívnej intuitívnej axiomatiky, vytvorenej na základe štúdia množstva konkrétnych špeciálnych prípadov a analógií, čo nie je nič iné, ako predstavy.

Znamená to teda, že máme celú nádhernú deduktívnu stavbu našej súčasnej matematiky zavrhnúť a vrátiť sa k Eulerovým predstavám? Pravdaže, nie. Okrem všetkých veľmi vážnych dôvodov, ktoré všetci dobre poznáme a ktorých podstatou je to, že matematika ako každá iná veda má svoj vnútorný život, ktorý sa prirodzene vyvíja a do ktorého nemožno robiť hrubé zásahy, je tu možno ešte dôvod, ktorý práve súvisí s aplikáciami. Dnes sa totiž matematika používa na modelovanie mnohých iných a nielen fyzikálnych dejov. Modely tu nie sú zďaleka také priliehavé,

ako modely fyzikálne, a preto ani overovanie výsledkov na základe predstáv nie je také presvedčivé.

Tak teda, čo je vektor? Šípka, prvok abstraktného vektorového priestoru, posunutie alebo trieda ekvivalencie? Myslím, že správna odpoveď je — všetko. Lingvisti hovoria, že koľko rečí vieš, toľkými životmi žiješ. Našu úvahu by sme mohli zakončiť voľnou parafrázou ich hesla: Čím bohatší register predstáv máš, tým lepšie veci rozumieš, a tým skôr niečo vymyslíš než ten, čo sa strnule drží jednej predstavy alebo formálnej definície.

Pavol Brunovský

Literatúra

1. Gedeonová, E.: Vektory očami algebraika. Matematické obzory, 8. Bratislava, ALFA 1975, s. 11 až 32.
2. Medek, V.: Vektory očami geometra. Matematické obzory, 9. Bratislava, ALFA 1976.
3. Franek, M.: Vektory očami stredoškolského profesora. Matematické obzory, 10. Bratislava, ALFA 1977, s. 1 až 18.
4. Zelinka, B.: Vektory očima funkcionálního analytika. Matematické obzory, 13. Bratislava, ALFA 1979.
5. Zelinka, B.: Vektory očima diskrétního matematika. Matematické obzory, 15. Bratislava, ALFA (v tlači).

P. Brunovský

Koniec chaosu?

Pokroky matematiky, fyziky a astronómie, 40(5) (1995), č. 5
233-243.

Koniec chaosu?

Pavel Brunovský, Bratislava

Právom by ste mi mohli vyčítať, že vodu kážem a víno pijem. Názvom článku som sa totiž prispôsobil móresom, ktoré v článku budem kritizovať. Jednoducho som ani ja nevedel odolať pokušeniu zneužiť to, že pôvodný význam slova chaos sa s jeho matematickým obsahom celkom nekryje. Na rozdiel od mnohých chaosomanov si to však dobre uvedomujem, kajam sa a sľubujem, že sa v ďalšom striktne obmedzím na matematický obsah pojmu.

Dva razy v ostatných dvoch desaťročiach vzišli z matematiky podnety, ktoré vzrušili širokú intelektuálnu verejnosť. Boli to teória katastrôf a chaos. Neviem, či to má historický precedens. Načim si teda položiť otázku, akú úlohu v tom zohrali ich názvy, ktoré navodzujú o veci neadekvátne predstavy. V tejto súvislosti si spomínam, ako sa slovenský preklad krásnej (pre matematikov) Arnoľdovej brožúrky „Teória katastrôf" dostal do predaja v bratislavských novinových stánkoch. Môj synovec, ktorý si ju pre jej názov kúpil, bol prekvapený najprv tým, že ma v nej našiel čo by recenzenta prekladu a potom aj jej obsahom.

Popri spoločnej popularite si však hodno povšimnúť fundamentálne odlišnosti katastrôf a chaosu.

Teória katastrôf je ucelená filozofická teória vychádzajúca z nemenej ucelenej matematickej teórie singularít. Snaží sa podľa jednotného princípu klasifikovať spontánny vznik foriem a náhle zmeny správania pri postupnej zmene parametrov bez ohľadu na ich materiálnu podstatu.

Takáto ambícia urobiť z teórie vševysvetľujúci princíp je v matematike ojedinelá, v minulosti sa skôr vyskytla u fyzikálnych teórií. Spoločné pre snahy tohoto druhu je, že ich propagátori, medzi nimi aj renomovaní vedci, často pri nich zachádzajú priďaleko a veľmi zľavujú z kritérií na vedeckosť.

Rozvírený prach okolo katastrôf medzičasom uľahol a teória sa dostala tam, kam patrí — medzi najkrajšie výsledky matematiky 20. storočia.

Na rozdiel od teórie katastrôf nie je chaos vlastne nijakou teóriou, ale pozorovaním, hoci prevratným. Nemá ani celkom jednotnú matematickú definíciu. Podstatou pozorovania, ktorá sa odráža vo všetkých jeho matematických definíciách, však je, že vďaka extrémnej citlivosti na počiatočné dáta a prepletenosti dynamiky sa systémy, ktoré sa v krátkom časovom úseku javia ako deterministické, dlhodobo môžu správať nepredvídateľne, eraticky.

Tomuto poznatku predchádzalo v dvadsiatych rokoch pozorovanie, ktoré bolo podľa mňa nemenej objavné. Vzišlo zo štúdie Van der Polovho lampového generátora

Prof. RNDr. PAVEL BRUNOVSKÝ, DrSc. (1934), Ústav aplikovanej matematiky, Univerzita Komenského, Mlynská Dolina, 842 15 Bratislava.

i Volterrovej-Lotkovej rovnice spoločenstva dravec-korisť a jeho podstatou je, že ustálené periodické cykly nie sú niečím patologickým, ale normálnym prírodným javom. Pokiaľ viem, tento poznatok sa s nejakou mimoriadnou pozornosťou nestretol.

História chaosu oslávila nedávno storočnicu. Prvé pozorovania o zložitosti dynamiky v okolí transverzálnej homoklinickej trajektórie (pozri Dodatok) pochádzajú — akože inak — od Poincarého [Po].

Je neveľmi známe, že Van der Polova rovnica zohrala významnú úlohu aj vo vývoji poznatkov o chaotickej dynamike. Odhalili ju v nej Cartwrightová a Littlewood pri štúdiu jej vynútených kmitov koncom štyridsiatych rokov (pozri [CL, L, Le]). Veľa pozornosti však ich výsledky vtedy nevzbudili. Mimo špecialistov nik nezaregistroval ani konštrukciu ďalšieho matematického objektu vytvárajúceho chaos začiatkom šesťdesiatych rokov — Smaleovej podkovy [S1, S2]. Šarkovského práce o komplikovanej dynamike spojitých zobrazení intervalu do seba [Š] zostali dosť neznámymi aj medzi špecialistami. Pritom exotické usporiadanie prirodzených čísel ($3 < 5 < 7 < \ldots < 2 \cdot 3 < 2 \cdot 5 < \ldots < \ldots < 2^k \cdot 3 < 2^k \cdot 5 < \ldots < \ldots < 2^3 < 2^2 < 2^1 < 1$), ktoré z neho vzišlo, by nemalo nechať chladným nikoho, kto v matematike nachádza kus dobrodružstva a krásy.

Väčší záujem už vzbudil Lorenzov zjednodušený model turbulencie ovzdušia [Lo]. Naozajstná explózia záujmu o chaos nastala, až keď ho Li a Yorke tak nazvali [LY]. Pritom ich práca, ktorá vznikla pri snahe pochopiť Lorenzov model, sčasti znovuobjavuje to, čo bolo oveľa presnejšie známe zo Šarkovského prác. Ťažko povedať, či tento prudký nárast záujmu bol naozaj vyvolaný šťastne zvoleným termínom, alebo je to len časová zhoda. V žiadnom prípade to však z mojej strany nemá byť kritika. Naopak, je užitočné sa poučiť, že aj pri rozširovaní vedeckých výsledkov treba myslieť nielen na obsah, ale aj na obal.

Či už to spôsobil názov alebo nie, práca [LY] značí začiatok dvoch desaťročí snáh odhaliť chaos, kde sa len dá — teoreticky, numericky a experimentálne.

Akú hodnotu majú výsledky týchto pokusov? Predovšetkým si treba pripomenúť, že chaos je matematický pojem a jeho prítomnosť možno dokázať iba matematickými prostriedkami. Vzhľadom na zaokrúhľovacie chyby a chyby metódy, ako aj na to, že na počítanie máme obmedzený čas, nie je napríklad numerickým sledovaním priebehu nejakého deja presne vzaté možné rozlíšiť, či má veľmi dlhú periódu, alebo je chaotický.

Nechcem tým zo štúdie chaosu vylúčiť numeriku. Napokon aj pri iných numerických štúdiách akceptujeme numerické výsledky, hoci nepotvrdzujú nejakú skutočnosť s úplnou presnosťou. Seriózne závery z numerických výsledkov však možno robiť iba vtedy, ak sú zdôvodniteľné v rozumne presnej miere.

A tu je v prípade chaosu kameň úrazu. Citlivá závislosť od počiatočných podmienok, ktorá je základným rysom chaosu, má za následok exponenciálny nárast dôsledkov nepresností počítania. Na prvý pohľad teda robí výsledky numerických výpočtov trajektórií chaotických systémov bezcennými.

O čo ide, vysvetlíme si na jednoduchom príklade dynamického systému

$$x_{t+1} = 2x_t. \tag{1}$$

Jeho trajektórie (t.j. postupnosti generované rekurentným predpisom (1)) sú geometrické postupnosti, ktorých všeobecný člen vieme presne spočítať — platí

$$x_t = 2^t x_0.$$

Predstavme si však, že by sme členy trajektórie počítali rekurentne priamo zo vzťahu (1), ako sa to robí v zložitejších prípadoch. Pri počítaní v dvojnásobnej presnosti na bežnom počítači sa môžeme pri každom výpočte dopustiť nepresnosti $\delta \sim 10^{14}$. To značí, že v prvom kroku môžeme namiesto hodnoty

$$x_1 = 2x_0$$

vypočítať hodnotu

$$\tilde{x}_1 = 2x_0 + \delta.$$

Ak by sme sa už nijakej inej chyby v ďalších výpočtoch nedopustili, po 50 krokoch výpočtu dostaneme člen

$$\tilde{x}_{50} = 2^{49}\tilde{x}_1 \geqq x_{50} + 10^{-14}2^{49} \geqq x_{50} + 1,$$

teda neúnosne veľkú chybu rádu jednotiek.

V tomto príklade samozrejme nijakého chaosu niet. Aby však systém mohol byť chaotický, musí obsahovať expanzívnu zložku, ktorá sa správa podobne (špecialistom pripomenieme, že to vyplýva napr. z toho, že aspoň jeden Ľapunovov exponent musí byť kladný).

Exponenciálny nárast dôsledkov numerických nepresností počítania trajektórie daného bodu je teda neodlúčiteľný od chaosu. Zdanlivo z toho vyplýva, že úspech pri numerickom pátraní po chaose automaticky spochybní metódu jeho hľadania. Hoci sa nedá celkom povedať, že by sa nik nad týmto paradoxom nebol zamyslel [LL], bezmála 15 rokov si väčšina ľudí, „počítajúcich" chaos z toho veľa starostí nerobila. Až nedávno sa našli ľudia, ktorí sa odvážili povedať, že „kráľ je nahý". Bol to napríklad numerik E. Adams z Ústavu aplikovanej matematiky v Karlsruhe.

Polemike okolo chaosu venoval časopis Spiegel v roku 1993 sériu troch dlhých článkov [B]. Podľa nich E. Adams zašiel až tak ďaleko, že spochybnil existenciu chaosu vo všeobecnosti a označil ho za numerický artefakt (treba však poznamenať, že Adams v odpovedi [Ad] na článok [RDP] uvádza, že v [B] boli jeho výroky prekrútené). Paradoxne sa to stalo v dobe, keď už kráľ niekoľko rokov nahý nebol a keď už bolo známe, že artefaktom je uvedený rozpor.

Záchrana je v tom, že pri počítaní chaosu sa sleduje náhodne vybraná trajektória. Nie je teda vôbec dôležité, aby bola trajektóriou vopred (zvyčajne aj tak náhodne) vybraného počiatočného bodu. Ukážeme si, že v príklade (1) numericky získaná trajektória $\{x_t\}$ zostáva v blízkosti nejakej trajektórie, prípadne iného počiatočného bodu než je $\tilde{x}_0 = x_0$.

Za tým účelom uvažujeme všeobecnejší neautonómny (t.j. od t závislý) lineárny rekurentný vzťah

$$x_{t+1} = a_t x_t, \tag{2}$$

kde $0 < |a_t| < a < 1$, a týmto vzťahom generované postupnosti (ktoré pre jednoduchosť nie celkom oprávnene budeme podobne ako v prípade autonómneho vzťahu (1) nazývať trajektóriami). Ľahko si overíme, že pre trajektóriu $\{x_t\}$ vzťahu (2) platí

$$x_t = a_{t-1} \dots a_\tau x_\tau \qquad \text{pre } t \geqq \tau, \tag{3}$$

$$x_t = a_t^{-1} \dots a_{\tau-1}^{-1} x_\tau \qquad \text{pre } t \leqq \tau. \tag{4}$$

Nech teraz $\{\tilde{x}_t\}$ je postupnosť, počítaná podľa (2) s nepresnosťou $\leqq \delta$, t.j. platí

$$\tilde{x}_{t+1} = a_t \tilde{x}_t + \delta_t, \tag{5}$$

kde

$$|\delta_t| \leqq \delta, \quad \delta > 0; \tag{6}$$

takúto postupnosť budeme v ďalšom nazývať δ-pseudotrajektóriou. Odčítaním (2) a (5) dostávame

$$|\tilde{x}_{t+1} - x_{t+1}| \leqq a|\tilde{x}_t - x_t| + \delta. \tag{7}$$

Ak pseudotrajektória $\{\tilde{x}_t\}$ a trajektória $\{x_\tau\}$ vychádzajú z toho istého bodu v okamihu $t = \tau$, t.j. platí

$$x_\tau = \tilde{x}_\tau, \tag{8}$$

sčítaním nerovnosti (7) od τ do $t-1$ dostávame pre $t \geqq \tau$

$$|\tilde{x}_t - x_t| \leqq (1 + a + \dots + a^{t-\tau-1})\delta \leqq \frac{1}{1-a}\,\delta. \tag{9}$$

Ak teda platí (3), pre $t \geqq \tau$ sa členy trajektórie a numericky počítanej δ-pseudotrajektórie budú odlišovať nanajvýš o $\varepsilon\ (= (1-a)^{-1}\delta)$, ktoré je pevným od dĺžky postupnosti nezávislým násobkom presnosti počítania δ.

Obrátením času (t.j. zámenou t na $-t$) sa môžeme jednoducho presvedčiť, že ak namiesto (3) platí

$$|a_t| > a > 1, \tag{10}$$

(čo spĺňa aj (1)), namiesto (8) dostávame pre δ-pseudotrajektóriu (2) spĺňajúcu (8) odhad

$$|\tilde{x}_t - x_t| \leqq \frac{1}{1-a^{-1}}\,\delta, \tag{11}$$

avšak pre $t \leqq \tau$. Teda napriek tomu, že v prípade (10) sa δ-pseudotrajektória vychádzajúca z bodu x_0 môže od presnej trajektórie tohoto bodu exponenciálne vzďaľovať, musí bez ohľadu na dĺžku pseudotrajektórie $\{\tilde{x}_\tau\}_{t=0}^{\tau}$ existovať presná trajektória $\{x_t\}_{t=0}^{\tau-1}$, ktorej členy sa od členov postupnosti $\{\tilde{x}_t\}_{t=0}^{\tau}$ líšia o nezávislé od dĺžky postupnosti číslo $\varepsilon\ (= (1-a_{-1})^{-1}\delta)$. Konkrétne je to trajektória systému (2), spĺňajúca $\tilde{x}_\tau = x_\tau$. Poznamenajme, že x_0 a $\tilde{x}_0$ sa zhodovať nemusia a že trajektória $\{x_t\}$ závisí od τ.

Jednoduchým limitným prechodom sa však možno presvedčiť, že trajektóriu $\{x_t\}$ spĺňajúcu (11) možno vybrať nezávisle od τ; takúto trajektóriu nazveme ε-tieňom pseudotrajektórie $\{\tilde{x}_t\}$.

Nič nám teraz nebráni rozšíriť pozorovanie o existencii ε-tieňa pre vektorový lineárny rekurentný predpis s dvoma zložkami typu (2), z ktorých jedna spĺňa (3) a druhá (10). Voľne povedané sa ε-tieň získa limitným prechodom riešenia okrajovej úlohy v ktorej zbiehavá zložka je rovná členu pseudotrajektórie na začiatku a rozbiehavá na konci zvoleného časového úseku. Ba dokonca ho možno rozšíriť na vektorový rekurentný predpis ľubovoľnej dimenzie

$$x_{t+1} = A_t x_t, \tag{12}$$

$x_t \in \mathbb{R}^n$, kde A_t sú regulárne matice typu

$$A_t = \begin{pmatrix} B_t & 0 \\ 0 & C_t \end{pmatrix}, \tag{13}$$

a teda (13) možno písať ako dvojicu vzťahov

$$\begin{aligned} y_{t+1} &= B_t y_t, \\ z_{t+1} &= C_t z_t \end{aligned} \tag{14}$$

s maticami B_t, C_t, ktorých rozmery nezávisia od t, spĺňajúcimi

$$|B_{\tau+t-1} \dots B_\tau y| \leqq K a^t |y|, \tag{15}$$

$$|C_{\tau+t-1} \dots C_\tau z| \geqq K^{-1} a^{-t} |z| \tag{16}$$

pre $t \geqq 0$ a $0 < a < 1$.

Všimnime si ďalej, že na existencii ε-tieňovej trajektórie, kde $\varepsilon \leqq Q\delta$ pre nejaké $Q > 0$, stačí, aby sa systém vzťahov (12) dal na podobu (13) pretransformovať postupnosťou lineárnych transformácií

$$(y_t, z_t)^{\mathrm{T}} = S_t x_t$$

($^{\mathrm{T}}$ označuje transpozíciu), pokiaľ $|S_t|$ a $|S_t^{-1}|$ sú rovnomerne ohraničené. V reči matíc žiadame na rozdiel od (13) od matíc A_t, aby spĺňali

$$A_t = S_{t+1}^{-1} \begin{pmatrix} B_t & 0 \\ 0 & C_t \end{pmatrix} S_t, \tag{17}$$

kde B_t a C_t spĺňajú (13). O systéme (12), ktorý má túto vlastnosť, hovoríme, že vykazuje exponenciálnu dichotómiu.

Zo zrejmých dôvodov možno prítomnosť nelineárnej expanzívnej zložky z v systéme rekurentných vzťahov (14) interpretovať ako citlivú závislosť na počiatočných dátach. Všeobecne sa však od chaosu okrem citlivej závislosti od počiatočných dát vyžaduje

ešte „tranzitívnosť“, teda existencia hustej trajektórie. Táto požiadavka vyjadruje, že sa systém nepredvídateľne môže ľubovoľne blízko priblížiť k ľubovoľnému stavu.

Takúto vlastnosť môžu mať iba nelineárne dynamické systémy. Treba teda pre ne nejako prispôsobiť pojem dichotómie trajektórií.

Uvažujme dynamický systém

$$x_{t+1} = f(x_t), \tag{18}$$

kde $x_t \in \mathbb{R}^n$ (alebo všeobecnejšie z n-rozmernej diferencovateľnej variety) a f je C^1-difeomorfizmus (teda spojité diferencovateľné zobrazenie s diferencovateľným inverzným).

Hovoríme, že invariantná množina H systému (18) je hyperbolická, ak pre ľubovoľnú trajektóriu $\{x_t\}$ systému (18) v H má linearizovaný systém (12) s $A_t := Df(x_t)$ exponenciálnu dichotómiu s rovnakými dimenziami y a z, rovnakými konštantami K a s rovnakým ohraničením C na transformačné matice S_t a S_t^{-1}.

Platí

Tieňová lema. *Nech H je kompaktná invariantná hyperbolická množina dynamického systému* (18). *Potom existuje $Q > 0$ také, že ľubovoľná δ-pseudotrajektória s dostatočne malým $\delta > 0$ má jediný ε-tieň, kde*

$$\varepsilon \leqq Q\delta. \tag{19}$$

Tieňová lema je objavom D. V. Anosova z roku 1963. Je súčasťou rozsiahlej práce [A], bohatej na originálne myšlienky a postupy. Jej dôkaz je technicky náročný.

Tieňová lema nám vlastne tvrdí to, čo treba — ak počítame trajektóriu numericky s presnosťou δ, bude v jej ε-blízkosti presná trajektória. Vzhľadom na odhad (19) môžeme teoreticky ε voliť ľubovoľne malé. Prakticky je však δ zdola ohraničené zaokrúhľovacou chybou počítača, a to nám dáva dolné ohraničenie na ε. Preto je zaujímavá aj konštanta Q, ktorá sa odvodzuje od konštánt exponenciálnej dichotómie K, a, C.

Horšie je to s predpokladmi hyperboličnosti, na ktorej overovanie nie sú vypracované všeobecne účinné metódy. Súvisí s prítomnosťou Ľapunovových exponentov rozličných znamienok, známym to heuristickým predpokladom chaosu.

Matematicky dokázateľne chaotické množiny sú všetky hyperbolické aspoň v zoslabenej forme (jeden z príkladov je v Dodatku, kde tieňová lema vystupuje ako prostriedok dokazovania chaosu). Nie je preto veľkým prehreškom, ak tieňovú lemu prijmeme ako zdôvodnenie korektnosti numerických výpočtov chaotických trajektórií.

Stretol som sa aj z názorom, že v pozadí rozruchu okolo Adamsových názorov sú aj peniaze. Podľa tohoto názoru nemeckej meteorologickej lobby domáhajúcej sa prostriedkov na stále výkonnejšie počítače neveľmi vyhovovalo, že sa počasie považovalo za typicky chaotický, a teda dlhodobo nepredpovedateľný jav. Veľmi jej teda prišlo vhod tvrdenie, že chaos je artefakt. Pre spravodlivosť treba povedať, že som toto počul od človeka, ktorého úspešná kariéra stála práve na „počítaní“ chaosu.

Nad zdôvodniteľnosťou numerickej simulácie chaotických procesov sa ako prvý zamyslel spoluautor jeho názvu J. Yorke so spolupracovníkmi [HYG1]. Myšlienka použitia tieňovej lemy patrí Chowovi a Palmerovi [CP1, CP2].

V spomínaných a ďalších prácach [HYG2, SY1, SY2, CP3, CV] sa autori nespoliehajú na vieru v ťažko overiteľný predpoklad rovnomernej hyperboličnosti chaotickej množiny. Nahradzujú ho buď a posteriori alebo rekurentne numericky overovanou dichotómiou linearizácie pozdĺž numericky počítanej individuálnej trajektórie. Umožňuje im to nielen rigorózne dokázať existenciu tieňovej trajektórie, ale aj odhadnúť jej vzdialenosť od numericky získanej pseudotrajektórie. Stojí za poznámku, že o týchto prácach niet v sérii článkov [B] ani zmienky.

Prirodzenou otázkou je, ako je to s dynamickými systémami so spojitým časom, generovanými diferenciálnymi rovnicami. Tieňová lema pre takéto systémy je sformulovaná a dokázaná v [KP], numerickým výpočtom je venovaná práca [CV].

Na záver ešte jedna dobrá správa pre fanúšikov chaosu: Po dvadsiatich rokoch neúspešných pokusov mnohých matematikov sa nedávno K. Mischaikowovi a M. Mrózekovi podarilo vypracovať dôkaz chaotičnosti atraktora v Lorenzových rovniciach. Využíva sa v ňom počítač, ale na celkom inej úrovni: pri precíznom sledovaní nepresnosti počítania sa overujú algebraicky — topologické podmienky toho, že atraktor obsahuje „Smaleovu podkovu", v ktorej chaotičnosť je dokázaná.*) Ani tento výsledok nevzbudil pozornosť autora [B], kde je chaotičnosť Lorenzovho atraktora spochybnená. Cituje sa v ňom (či už oprávnene alebo nie) výrok E. Adamsa, podľa ktorého „Obrázok (myslí sa zmeny počítačom získaný obrázok trajektórie v atraktore Lorenzových rovníc) je takmer celkom nesprávny. Čo ukazuje, nemá nič spoločného s tým, čo sa v Lorenzovej rovnici odohráva". Sérii článkov Spieglu však treba priznať jeho oprávnenú kritiku nekvalifikovaného zbožňovania chaosu. Hoci to takto explicitne neformuluje, autor cíti neprimeranosť stotožňovania rozličných významov pojmu.

Ako teda odpovedať na otázku v nadpise? Milovníci chaosu, ktorým Adamsove vývody spôsobili bezsenné noci, môžu naďalej pokojne spávať — pokiaľ im k tomu postačí, že ich výpočty nie sú ich úspechom automaticky vylúčené. Celkom bezstarostní však zasa byť nemôžu. Existuje totiž príklad, v ktorom je naozaj možné chaos v nechaotickom systéme „vyrobiť" nepremyslenou numerickou aproximáciou (pozri Dodatok).

Dodatok

Triviálnym prípadom hyperbolickej množiny dynamického systému (18) je hyperbolický pevný bod. Je to bod $\hat{x}$ taký, že

$$f(\hat{x}) = \hat{x}$$

*) Presnejšie povedané, atraktor Lorenzových rovníc obsahuje invariantnú podmnožinu, tok na ktorej je semikonjugovaný Smaleovej podkove.

a vlastné hodnoty $Df(\hat{x})$ sú mimo jednotkovej kružnice. Vo vhodne volených súradniciach totiž vtedy možno písať

$$Df(\hat{x}) = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix},$$

kde vlastné hodnoty matice B majú absolútnu hodnotu < 1, kým absolútne hodnoty vlastných hodnôt C sú > 1 a B, C sú v Jordanovom kanonickom tvare. Priamočiarym výpočtom sa možno presvedčiť, že systém (12) s $A_t \equiv Df(\hat{x})$ má exponenciálnu dichotómiu, odhady (15), (16) sa zredukujú na

$$|B^t y| \geqq K a^t |y|, \tag{19}$$

$$|C^t z| \geqq K^{-1} a^{-t} |z| \tag{20}$$

$(0 < a < 1)$ pre $t \geqq 0$.

Odhady (19), (20) značia, že násobením maticou $Df(\hat{x})$ sa body invariantného podpriestoru $z = 0$ k sebe (a teda aj k 0) exponenciálne približujú, kým body priestoru $y = 0$ sa od seba exponenciálne vzďaľujú. Preto nazývame priestor $z = 0$ stabilným a priestor $y = 0$ nestabilným podpriestorom bodu $\hat{x}$.

Veta o stabilnej a nestabilnej variete hovorí, že existujú invariantné variety $W^s(\hat{x})$ (stabilná) a $W^u(\hat{x})$ (nestabilná) systému (18), prechádzajúce bodom $\hat{x}$ a dotýkajúce sa v ňom stabilného, resp. nestabilného podpriestoru. Tieto variety dedia vyššie spomínané vlastnosti približovania, resp. vzďaľovania sa trajektórií linearizovaného systému. Môžu sa pretínať aj v inom bode x_0 než v $\hat{x}$. Z ich invariantnosti vyplýva, že vtedy spolu s bodom x_0 obsahujú aj celú jeho trajektóriu $\{x_t\}, x_t = f^t(x_0)$ a že platí

$$\lim_{t \to \pm\infty} x_t = \hat{x}. \tag{21}$$

Trajektóriu $\{x_t\}$ nazývame homoklinickou trajektóriou bodu $\hat{x}$. Hovoríme že $\{x_t\}$ je transverzálna, ak dotykové priestory $T_{x_0}W^s(\hat{x})$, $T_{x_0}W^u(\hat{x})$ variet $W^s(\hat{x}), W^u(\hat{x})$ v bode x_0 majú jednobodový prienik (a teda ich algebraický súčet je celý priestor).

Ak $\{x_t\}$ je transverzálna homoklinická trajektória hyperbolického pevného bodu $\hat{x}$, potom množina

$$H := \{\hat{x}\} \cup \{x_t\}_{t=-\infty}^{\infty}$$

je kompaktná invariantná hyperbolická množina [P].

Aby sme si to ozrejmili, všimnime si že H sa skladá z dvoch trajektórií. Hyperboličnosť $\{\hat{x}\}$ značí exponenciálnu dichotómiu tejto jednobodovej trajektórie, zostáva teda ozrejmiť si exponenciálnu dichotómiu trajektórie $\{x_t\}$.

Linearizovaný systém (12) s $A_t = Df(x_t)$ zrejme zobrazuje dotykové priestory invariantných variet v bode x_t na dotykové priestory v bode x_{t+1}. Nie je ťažko si predstaviť, že dotykové priestory $T_{x_t}W^s(\hat{x})$ a $T_{x_t}W^u(\hat{x})$ sa pre $t \to \pm\infty$ blížia k stabilnému, resp. nestabilnému priestoru v bode $\hat{x}$ a preto linearizovaný systém (12) body na nich asymptoticky exponenciálne približuje, resp. vzďaľuje. Od t závislá lineárne

transformácia, ktorá prevádza $T_{x_t}W^s(\hat{x})$, resp. $T_{x_t}W^u(\hat{x})$ do podpriestorov $y = 0$, resp. $z = 0$ teda prevádza linearizovaný proces do tvaru, v ktorom platia odhady (15), (16).

Na invariantnej množine H teda platí Tieňová lema. Použijeme ju pre špeciálne pseudotrajektórie.

Zvolíme malé $\delta > 0$. Keďže platí (21), existuje N tak veľké, že x_{-N} a x_N sú v δ-okolí bodu $\hat{x}$. Označíme

$$\Sigma = \{x_{-N}, \ldots, x_0, \ldots x_N\}$$

a pre $m \geqq 0$ označíme Γ_m m-člennú postupnosť, ktorej všetky členy sú $\hat{x}$. Pre ľubovoľne zvolenú postupnosť nezáporných čísel $\{m_k\}_{k=-\infty}^{\infty}$ je postupnosť

$$\ldots, \Gamma_{m_{-1}}, \Sigma, \Gamma_{m_0}, \Sigma, \Gamma_{m_1}, \ldots,$$

ktorá vznikne zreťazením konečných postupností Γ_{m_k} a medzi nich vložených konečných postupností Σ zrejme δ-pseudotrajektóriou. Existuje teda jej ε-tieň. Ak δ (a teda ε) je dosť malé, potom jeho úsek, zodpovedajúci Σ predstavuje jeden obeh v blízkosti homoklinickej trajektórie, kým v úseku, zodpovedajúcom Γ_m, tieňová trajektória zostáva v blízkosti pevného bodu $\hat{x}$. Voľbou čísel m_k (pripomínam, že m_k môže byť aj 0) si teda pre tieňovú trajektóriu môžeme predpísať ľubovoľné počty následných obehov v blízkosti homoklinickej trajektórie a pomedzi ne vložiť pobyty trajektórie v blízkosti $\hat{x}$ ľubovoľnej dĺžky. Vidno teda, že správanie sa trajektórie je naprosto nepredvídateľné, eratické.

Spresnením našich argumentov možno dokázať, že dynamický systém v okolí homoklinickej trajektórie obsahuje invariantnú množinu, na ktorej je konjugovaný s tzv. Bernoulliho posunom, o ktorom je známe, že má všetky bežne prijímané vlastnosti chaosu: citlivú závislosť na počiatočných dátach, hustú trajektóriu, periodické body ľubovoľnej periódy, ergodickú invariantnú mieru, atď. Štúdium transverzálnej homoklinickej trajektórie bolo aj motiváciou pre vznik Smaleovej podkovy — zobrazenia roviny do seba, ktoré bolo jednou z prvých príkladov chaosu.

Príklad transverzálnej homoklinickej trajektórie naznačuje aj to, ako možno chaos vyrobiť umelo: homoklinická trajektória dynamického systému so spojitým časom nemôže byť transverzálna a ani nemá nič spoločné s chaosom. Diskretizáciou, resp. numerickou aproximáciou rovnice možno transverzálnu trajektóriu a teda aj chaos vyrobiť. Našťastie, miera transverzality klesá s veľkosťou kroku diskretizácie rýchlejšie, ako ľubovoľná jeho mocnina. Preto je v tomto prípade možné vyhnúť sa umelému chaosu voľbou dostatočne jemného kroku [FS].

Záverom by som sa rád poďakoval anonymnému recenzentovi, ktorý ma upozornil na všeličo, o čom som nevedel — napríklad na knihu [LL], ako aj na dozvuky série [B] v článkoch [RDP] a [Ad].

Literatura

[A] D. V. Anosov: *Geodetičeskije potoki na zamknutych rimannovych poverchnosťach otricatel'noj krivizny.* Trudy Mat. Inst. im. Steklova *90* (1967).

[Ad] E. ADAMS: Phys. B1 *50* (1994), 359.

[B] P. BRÜGGE: *Der Kult um das Chaos.* Der Spiegel *47* (1993); č. 39, 156–164; č. 40, 232–241; č. 41, 240–252.

[CL] M. L. CARTWRIGHT and J. E. LITTLEWOOD: *On nonlinear differential equations of the second order. The equation* $\ddot{y} + k(1 - y^2)\dot{y} + y = b\lambda k \cos(\lambda t + a)$, *k large.* J. Lond. Math. Soc. *20* (1945), 180–189.

[CP1] S. N. CHOW and K. J. PALMER: *The accuracy of numerically computed orbits of dynamical systems.* In „Equadiff 7", proceedings of the Conference held in Prague 1989 (1990), Teubner, Leipzig, 74–76.

[CP2] S.-N. CHOW and K. J. PALMER: *On the numerical computation of orbits of dynamical systems: The one-dimensional case.* J. Dynamics and Differential Equations *3* (1991), 361–379.

[CP3] S.-N. CHOW and K. J. PALMER: *On the numerical computation of orbits of dynamical systems: the higher dimensional case.* J. Complexity *8* (1992), 398–423.

[CV] S.-N. CHOW and E. S. VAN VLECK: *A shadowing lemma approach to global error analysis for initial value ODEs.* SIAM J. Sci. Comput. *15*, No. 4, July 1994, 959–976.

[FS] B. FIEDLER, J. SCHEURLE: *Discretization of homoclinic orbits, rapid forcing and „invisible" chaos.* Preprint SC 91-5 (1991), Konrad-Zuse Zentrum für Informationstechnik Berlin, vyjde v sérii Memoirs of American Mathematical Society.

[HYG1] S. M. HAMMEL, J. A. YORKE and C. GREBOGI: *Do numerical orbits of chaotic dynamical processes represent true orbits?* J. Complexity *3* (1987), 136–145.

[HYG2] S. M. HAMMEL, J. A. YORKE and C. GREBOGI: *Numerical orbits of chaotic processes represent true orbits.* Bull. Amer. Math. Soc. *19* (1988), 465–470.

[L] N. LEVINSON: *A second order differential equation with singular solutions.* Ann. Math. *50* (1949), 127–153.

[Le] M. LEVI: *Qualitative analysis of the periodically forced relaxation oscillations.* Mem. AMS *214* (1981), 1–147.

[Lo] E. N. LORENZ: *Deterministic non-periodic flow.* J. Atmos. Sci. *20* (1963), 130 až 141.

[LL] A. J. LICHTENBERG, M. A. LIEBERMAN: *Regular and Statistical Motion.* Springer, Berlin 1983, 276.

[LY] T. Y. LI and J. A. YORKE: *Period three implies chaos.* Amer. Math. Monthly *82* (1975) 985–992.

[MM] K. MISCHAIKOW, M. MRÓZEK: *Chaos in the Lorenz equations: a computer assisted proof.* Center for Dynamical Systems & Nonlinear Studies, Georgia Inst. of Technology, Report Nr. CDSN93-123, 1993.

[P] K. J. PALMER: *Exponential dichotomos; the shadowing lemma and transversal homoclinic points.* Dynamics Reported *1* (1988), 265–306.

[Po] H. POINCARÉ: *Sur les équations de la dynamique et les problemes des trois corpes.* Acta Math. *13* (1890), 1–270.

[RDP] P. H. RICHTER, H. DULLIN, H.-O. PEITGEN: *Der Spiegel, das Chaos — und die Wahrheit.* Phys. B1 *50* (1994), 335.

[S1] S. SMALE: *Diffeomorphisms with many periodic points.* In: Differential and Combinatorial Topology (ed. S. S. Cairns), Princeton University Press, Princeton 1963, 63–80.

[S2] S. SMALE: *Differentiable dynamical systems.* Bull. Amer. Math. Soc. *73* (1967), 747–817.

[SY1] T. SAUER and J. A. YORKE: *Shadowing trajectories of dynamical systems.* In „Computer-Aided Proofs in Analysis" (eds. K. Meyer and D. Schmidt), Springer-Verlag, Berlin 1990, 229–234.

[SY2] T. SAUER and J. A. YORKE: *Rigorous verification of trajectories for the computer simulation of dynamical systems.* Nonlinearity *4* (1991), 961–979.

[Š] A. N. ŠARKOVSKIJ: *Koexistencija ciklov nepreryvnogo otobraženia prjamoj na sebja.* Ukrainskij matematičeskij žurnal *16* (1964), 61–71.

O Fermatových číslech

Michal Křížek, Praha

1. Úvod

Francouzský matematik Pierre Fermat (1601–1665) se proslavil nejen svou velkou a malou větou Fermatovou, ale také hypotézou, že všechna čísla tvaru

$$F_m = 2^{2^m} + 1 \quad \text{pro } m = 0, 1, 2, \ldots \tag{1}$$

jsou prvočísla. Ani jedno z těchto tří tvrzení však pravděpodobně nedokázal. Přiznával ale, že s důkazem domněnky o prvočíselnosti F_m si neví rady. Čísla F_m se po něm nazývají Fermatova čísla.

Pokud je F_m prvočíslo, říkáme, že je Fermatovým prvočíslem. Prvních pět členů posloupnosti (1), tj.

$$F_0 = 3, \quad F_1 = 5, \quad F_2 = 17, \quad F_3 = 257, \quad F_4 = 65537, \tag{2}$$

jsou prvočísla. K tomu, aby číslo $2^n + 1$ pro n přirozené bylo prvočíslem, je nutné, aby byl exponent n tvaru 2^m pro $m \in \{0, 1, 2, \ldots\}$. Je-li totiž k přirozené a $l \geqq 3$ liché, pak

$$2^{kl} + 1 = (2^k + 1)(2^{k(l-1)} - 2^{k(l-2)} + \ldots - 2^k + 1). \tag{3}$$

Odtud plyne, že číslo $2^n + 1$ je složené, jestliže je exponent n dělitelný lichým číslem $l \geqq 3$. To však v posloupnosti (1) nenastane.

RNDr. MICHAL KŘÍŽEK, DrSc. (1952), je pracovníkem Matematického ústavu AV ČR, Žitná 25, 11 567 Praha 1 (e-mail: krizek@earn.cvut.cz). Tato práce byla částečně podpořena grantem č. 201/94/1067 GA ČR.

P. Brunovský

Jediné na svete?

Nové slovo, 5.12.1974.

Spolupráca matematikov socialistických krajín

Jedinė na svete

Koncom septembra navštívil Bratislavu riaditeľ Medzinárodného matematického centra S. Banacha vo Varšave prof. Czeslaw Olech, ktorý tu **prednášal** na letnej škole **„DIFFORD—74"** zorganizovanej **Matematickým** ústavom **SAV. Prof.** Olech je členom korešpondentom Poľskej akadémie vied a popredným svetovým odborníkom v matematickej teórii optimálneho riadenia. Využili sme túto príležitosť, aby sme mu položili niekoľko otázok o zameraní, práci a plánoch Centra, ktoré je spoločnou ustanovizňou viacerých socialistických krajín.

Ako vzniklo Centrum a čo je jeho cieľom?

Centrum vzniklo v rámci dohody o mnohostrannej vedeckej spolupráci akadémií vied socialistických krajín v roku 1971. Dohoda o zriadení bola podpísaná 13. 1. 1972. Jej signatármi boli akadémie vied Bulharska, ČSSR, NDR, Poľska, Rumunska a ZSSR. Svoju činnosť začalo Banachovo centrum v januári 1973. Pomenované je po významnom poľskom matematikovi prvej polovice tohoto storočia. Vytvorenie centra nebolo ľahkou záležitosťou, pretože ona je dnes svetovým unikátom a nebolo preň vzorov.

Hlavnou myšlienkou, ktorá viedla k vytvoreniu Centra bola potreba prechodu na nové, efektívnejšie formy medzinárodnej spolupráce, v oblasti matematiky, najmä medzi socialistickými krajinami.

Aké zameranie a poslanie má spolupráca matematikov?

Základnou formou činnosti centra **sú** špecializované semestre. **V** danom časovom úseku, ktorý **sa obvykle kryje s** akademickým **semestrom, sa** činnosť Centra **sústreďuje okolo** vybranej matematickej disciplíny. Do Centra pozývame významných odborníkov, ako aj mladších matematikov — stážistov, **ktorí spoločne bádajú, prezentujú a** diskutujú **o** svojich vedeckých problémoch a výsledkoch. Na konci semestra sa obvykle usporiada medzinárodná konferencia na príslušnú tému.

Účastníkmi Centra sú „stážisti" a „prednášatelia": Prvými rozumieme mladých matematikov, ktorí sa zúčastňujú na celom semestri. Do druhej skupiny zahrňujeme špecialistov, ktorí prichádzajú do Centra na kratšiu alebo dlhšiu dobu, aby prednášali a viedli semináre. Treba poznamenať, že hranica medzi prvou a druhou skupinou nie je ostrá; mali sme už v Centre viacerých stážistov, ktorí dosahovali takú úroveň, že sami napr. boli schopní viesť niektorý seminár. Väčšina účastníkov je zo socialistických krajín, ale na práci Centra sa zúčastňujú aj matematici z iných krajín.

Akú organizačnú štruktúru má Centrum a ako sa určuje jeho program?

Centrum vedie vedecká rada, pozostávajúca z dvoch pracovníkov z každej signatárskej krajiny (z ČSSR akad. J. Novák, riaditeľ Matematického ústavu ČSAV a doc. L. Mišík, DrSc., zástupca riaditeľa Mat. ústavu SAV) a riaditeľ Centra. Vedecká rada vyberá témy semestrov, menuje ich organizačné výbory atď.

Ako je to s financiami?

Účastníci **zo** socialistických krajín **prichádzajú do** Centra na **svoje náklady; administratívne náklady hradí** Poľská akadémia vied. **Táto forma je** z administratívnej **stránky veľmi jednoduchá a umožnila urýchlene vytvoriť Centrum. Treba poznamenať, že hoci na Západe** existujú matematické inštitúcie s medzinárodnou pôsobnosťou, takáto inštitúcia dosiaľ neexistovala v socialistických krajinách a je aj prvou inštitúciou na svete, ktorá vznikla medzinárodnou dohodou a je riadená medzinárodným orgánom.

Mohli by ste zhrnúť doterajšie skúsenosti?

Môžem povedať, **že** Centrum **sa** znamenite osvedčilo. Konali sa už tri semestre, a to zo základov matematiky a matematickej logiky (jar 1973), z matematickej teórie optimálneho riadenia (jeseň 1973), z matematických základov informatiky (jar 1973) a teraz prebieha semester z globálnej analýzy.

K semestrom uvediem niekoľko čísel: Na prvom semestri sa zúčastnilo 35 stážistov **a** 22 takých prednášateľov, ktorí strávili v Centre viac ako mesiac. Viacero špecialistov strávilo v Centre kratšie časové úseky — 1 — 2 týždne. Vcelku sa na semestri zúčastnilo 132 osôb, na 600 hodinách vedeckých seminárov, prednášok atď. Účastníci mali k dispozícii 5 druhov skrípt. Počas semestra vzniklo dvadsaťpäť pôvodných vedeckých prác, z ktorých niektoré boli vytvorené kolektívom autorov z rôznych krajín (spoluautorom jednej z nich bol aj Bratislavčan dr. I. Korec, CSc.). Čísla z ostatných semestrov sú podobné.

Počet prác, ktoré vznikli priamo počas **semestra zďaleka** nedáva **plný obraz o jeho význame** a prínose. **Nie je** možné spočítať, **koľko myšlienok a nápadov, ktoré vznikli počas** semestra **sa realizuje až po jeho skončení. Sám som bol účastníkom jedného semestra (z matematickej teórie optimálneho riadenia) a mal som možnosť** využívať jeho pracovnú atmosféru. Hoci moje administratívne záväzky mi nedovolili zúčastniť sa na ňom v takej miere, ako by som si bol želal, najmenej tri moje práce by bez neho neboli vznikli.

Dojmy účastníkov **sú vo všeobecnosti** kladné, a často priam nadšené. Jeden z účastníkov mi povedal, že nikdy predtým tak intenzívne nepracoval a že si ani nepredstavoval, že je možné tak intenzívne pracovať.

Čo možno tu očakávať v budúcnosti?

Sú pripravené už 3 ďalšie semestre: numerické metódy a matematické modely (jar 1975), teória aproximácie (jeseň 1975) a teória pravdepodobností a stochastických procesov (jar 76). V budúcnosti chceme organizovať okrem semestrov aj akcie iného druhu (konferencie, sympóziá a i.)

Ako sa zúčastňujú matematici z Československa a Slovenska na práci Centra?

Československo od začiatku zaujalo k Centru kladný a konštruktívny postoj a tým sa nemálo pričinilo o jeho vznik. Prednášatelia i stážisti z ČSSR, medzi ktorými boli viacerí zo Slovenska, sa uviedli veľmi dobre a ich práce boli pre Centrum prínosom. V zrovnaní s počtami účastníkov napr. z NDR či Maďarska je ich však pomerne málo, čo bolo vzhľadom na úroveň čs. matematiky v niektorých disciplínach škodou pre ústav a domnievam sa, i pre čs. matematiku. Tu však čaká hádam na vyriešenie niektorých otvorených finančných otázok vysielania pracovníkov do Centra.

Zhováral sa PAVOL BRUNOVSKÝ

P. Brunovský

Konspekt príspevku: Rozmýšľanie o tom, kto je talent a načo ho vyhľadávať alebo, čo keby sa u nás narodil druhý Gauss ...

Nové slovo, 1984.

Nové Zámky 1992

1

Rozmýšľanie o tom, kto je to talent a načo ho vyhľadávať
alebo
čo, keby sa u nás narodil druhý Gauss ...

Motto:
Potym povedám, chlapci, my sme malý národ;
ak budeme ešte aj sprostí, bude s nami zle ...
J. Hronec

Veľa sa v ostatnom čase hovorí o vyhľadávaní vedeckých talentov a starostlivosti o ich rozvoj. Akosi sa pritom zabúda na oveľa základnejšie otázky: kto je to vlastne talent a prečo ho vlastne treba vyhľadávať a rozvíjať. Že je to jasné a nieto tu o čom pochybovať? Nuž, počkajme radšej s odpoveďou chvíľu ...

Začnime najprv ~~s~~ kúskom histórie. Slovenská matematika - ako aj slovenská veda vôbec, je mladá. Veď odhliadnuc od ~~mm~~ jednotlivých osobností sa systematicky začala rozvíjať vlastne až po oslobodení. Za tento čas sa urobil skutočne pokrok. Zvyčajne sa vyjadruje číslami: máme Akadémiu, ktorá má toľko a zoľko ústavov a zamestnancov, máme toľko a toľko kandidátov a doktorov vied, atď. Pozrime sa však na vec trocha z inej stránky: zodpovedá to, čo Slováci dali svetovej vede, oným počtom? Vie sa o nás vo svete? Podarilo sa nám vychovať vedcov svetového mena, ktorí vo vede urobili "dieru do sveta"?

Nevidím natoľko do iných vied, tak sa vrátim k matematike. Aby sme mali akú-takú mierku, porovnávajme so susedom najbližším-s českou matematikou. Nik nemôže byť prekvapený, že sme v minulosti nemali takú osobnosť, ako boli ~~Eduard Čech~~ Bernard Bolzano, alebo Eduard Čech. Ale že sa nám za bezmála 40 rokov nepodarilo vychovať ~~mladíka~~ mladíka, ktorý by bol urobil takú dieru do matematického sveta ako nedávna urobili Pudlák s Tůmom, to už stojí za zamyslenie.

V génoch to asi nebude - veď Pudlák je vlastne Slovák, ibaže vyrástol v Prahe. Keď to nie je v génoch, ostávajú už len podmienky. Poďme teda sledovať náš talent, ktorého kdesi na základnej, či strednej škole obetaví učitelia, vedúci krúžkov, či iní dobrovoľníci objavili, ~~zapálili~~ zažali v ňom

iskru nadšenia pre vedu a on sa teraz plný elánu a sveľkými cieľmi zapísal na štúdium ~~matematiky~~. Aby bolo hneď na začiatku jasné: nejde mi teraz o ~~taký~~ priemerný talent, ale o takého malého potenciálneho Gaussa. Veď prečo by sa raz nejaký Gausík nemohol narodiť aj na Slovensku? A nebola by vtedy našou povinnosťou voči svetu zabezpečiť mu také podmienky, aby svetu dal to, čo mu je schopný dať?

Vôbec, ako to bolo s Gaussom? ~~Ako xxxxxxxxxxxxxxxx~~ V nejakej veľmi nízkej triede jeho učiteľ usúdil, že ho už nemá čo naučiť, a poslal ho do škôl. Ujal sa ho vojvodca z Braunschweigu, ktorý financoval jeho ďalšej štúdiá - a to aj vtedy, keď si musel na to požičať. Keď mal voľačo cez 20 rokov, postavili mu observatórium, kde v mieri a pokoji pracoval pocelý svoj dlhý plodný život. A odplatil sa Gauss spoločnosti za to? Že či. Veď dal veľa nielen matematike a fyzike, ale napríklad aj takej praktickej veci ako je kartografia a okrem iného zmapoval Nemecko.

Aby bolo jasné, toto nemá byť óda na "staré dobré časy". Mohli by sme totiž s ľahkosťou nájsť aj opačné príklady. Veď napríklad taký Niels Abel (vynikajúci nórsky matematik) zomrel mladý v Paríži v podstate od biedy... Chcem tým povedať niečo iné: že naša spoločnosť ~~je tak vyspelá~~, že Gausík by v nej mal zákonite mať také podmeinky, ako ich Gauss vďaka osvietenému vojvodcovi mal náhodou ...

Vráťme sa teda ku Gausíkovi. Zapísal sa a študuje. Má učiteľov lepších i horších, ako to už býva. Lenže, keď je Gausík, nestačí mu to, čo dostane od učiteľov, a skoro sa chce postaviť na vlastné nohy. Na to potrebuje prinajmenej dve veci: čas a literatúru. S časom by to ešte hádam aj šlo - aj keď nemožno povedať, že by na internátoch vládla veľmi študijné ovzdušie a gausík, negausík - odchôdzovať si svoje musí. S literatúrou je to horšie - na Slovensku totiž prakticky neexistuje matematická knižnica. ~~xxxxxxxxxx~~ Mimochodom, to ešte Gausík nič nevie o perspektívach (objem devízových prostriedkov na nákup literatúry na r. 1984 = 20% objemu z r. 1983).

Tak, či onak - Gausík skončil. Pretože všetky sily venoval zdolávaniu chrámu vedy, pozabudol trocha na to, že sa treba venovať starosti o chlieb. Možno tajne dúfal, že jeho kvality budú dostatočné na to, aby ~~s týmto nemal problémy~~ mu otvorali dvere doktorát. Vy-

- 3 -

soké školy a Akadémia, kam by patril, majú stop stavy - ale, možno, že si jeho kvality predsa len niekto povšimne a Gausík sa načas uchytí na zástupovanie, nejakej devy, čo je práve na materskej.

S literatúrou sú naďalej problémy, ale Gausík už prišiel na to, že trocha sa tomu dá odpomôcť. Stojí to voľačo času, ale dá sa pozrieť referatívny časopis, niekde pozháňať autorovu adresu a napísať mu o separát. A nedávno to ešte aj tak bolo, že autor zvyčajne separát poslal a pošta ho Gausíkovi doručila. Lenže, aj v tom sme dnes ďalej. Miesto separátu príde Gausíkovi zelený lístok z colnice a nechcete odo mňa, aby som detailne opísal kalváriu, ktorú Gausík musí postúpiť, aby separát z colnice vydoloval - prípadne aj nevydoloval.

Dúfajme, že ho ani to ešte neotrávilo a Gausík nastúpil do ašpirantúry. A pretože je Gausík, pokračuje rýchlo.- možno, že by ju stihol aj za dva roky. Teda mu naložme: nech vyberá príspevky na Spolok priateľov žehu a nech zorganizuje ping-pongový turnaj a návštevu Cirkusu Humberto. Veď on tú ašpirantúru aj tak urobí v termíne.

Aj sa stalo - ašpirantúru urobil, a nie hocijako. Takže si možno jeho prácu povšimli vo svete. Dostáva žiadosti o separáty, čo má za následok pravidelné platby colnici ~~zvýšenie počtu návštev na colnici za účelom~~ ich odosielania. Ba dokonca sa možno aj stane, že ho niekam pozvú, aby o svojich výsledkoch prednášal. To by mu hádam malo zaistiť pozornosť aj doma. Veruže zaistí - začne byť podozrievaný zo spolkov s diablom. Veď kto to kedy slýchal, aby jeden Slováčisko čosi súceho vytvoril. To nemôže byť celkom po kostolnom poriadku.

Možno, že Gausík predsa len dôjde uznania aj doma. Observatórium mu síce nepostavia, ale mu možno zveria vedenie nejakej výskumnej úlohy, či oddelenia alebo katedry. A to je jeho definitívny koniec - lebo odvtedy celý svoj ďalší život strávi vysedávaním na poradách a písaním hlásení, prognóz, koncepcií, a iných podobne konštruktívnych činností.

Keď si to tak spočítame, vyjde nám, že pri dobrej vôli je šanca nanajvýš 1: 10, že z Gausíka bude nejaký osoh. Skôr z neho vyrastie ufrfľaný kverulant, s ktorým spoločnosť bude mať iba problémy.

- 4 -

Čo teda robiť? Predovšetkým musíme prehodnotiť naše predstavy o tom, kto je talent. Vezmime si takého mladíka, ktorý má akurát taký talent, aby mu škola nerobila problémy, ale v živote"to vie". Keďže mu pre uspokojenie jeho túžby po vedení stačí látka na skúšky, ostane mu dosť času na to, aby sa venoval svojej budúcnosti. Už za štúdia si našiel správneho patróna a kontakty, ktoré má vrodené, mu umožnia získať miesto aj tam, kde ho niet. Literatúra mu nechýba - veď mu stačí rozpracovávať to, čo už pred ním trikrať prežuli generácie vedátorov pred ním. A čo ak mu ide už na šiesty rok ašpirantúry a práca nikde? Treba mu dať všemožné úľavy, aby konečne miečo vypotil - veď Gausík to za neho odtiahne a ašpirantúru urobí i tak ľavou rukou ...

A máme tu jedného šarmantného vedátora, s ktorým nie sú nijaké problémy, a čo je hlavné - číslo v kolonke " vedeckí pracovníci" je o jedno vyššie. Že po ňom nič nezostane? Ale to nič, veď s tým sa v skutočnosti aj tak neráta ...

Takže kto je talent, sme už vyriešili. Zostáva otázka, čo ak objavíme potenciálneho Gausíka. Myslím, že odpoveď nájdeme v Elamovi Ohnivákovi, ktorý po svojich zlatokopeckých skúsenostiach našiel na svojom pozemku zlatú žilu, rýchlo ju zakopal, aby ju nik nenašiel a bol pokoj. A to urobme aj my: veďme Gausíka k tomu, aby hral na gitare, alebo hral basketbal, aj keď na to nemá najmenšie predpoklady. Bude to lepšie pre neho, aj pre spoločnosť ...

P. Brunovský

Je načase položiť latku vyššie

Nové slovo 29, 1987.

Nové Slovo 1984

Doktori vied a stratégia urýchlenia

Je načase položiť latku vyššie

Prof. Ebringer sa vo svojom článku zamýšľa nad praxou udeľovania titulov doktorov vied. Navrhuje spôsoby, ako dať do súladu túto prax s textom vyhlášky. Je však ovela jednoduchšie riešenie: prispôsobiť text vyhlášky praxi. Napríklad: Doktorom vied sa stáva osoba, dlhé roky pracujúca vo výskumnom pracovisku, ktorej doktorská dizertačná práca nie je o nič horšia ako iné práce úspešne obhájené a treba podporiť jej autoritu, lebo zastáva dôležité spoločenské postavenie.

Samozrejme, že to nemyslím vážne, ale celkom zasa nežartujem. S malými obmenami som totiž toto zdôvodnenie počul neraz, keď išlo o to, prečo práve v tomto prípade treba, v porovnaní s vyhláškou urobiť výnimku (samozrejme, vždy to bol prípad posledný).

Hovorí sa, že výnimka potvrdzuje pravidlo. Ako protiargument by som uviedol elementárny fakt z matematickej logiky: Ak v jednom tvrdení pripustíme rozpor, potom už je pravdivé čokoľvek. Poučenie, ktoré si z toho treba vziať je: Stačí zopár výnimiek a každý, komu ju neurobím, sa cíti ukrivdený. Navyše, výnimky sa šíria ako zhubný nádor: „výnimoční" doktori vied ako členovia komisií a oponenti majú tendenciu robiť ich ďalej, takže sa postupne stratí cit pre to, čo vlastne je pravidlom a čo výnimkou.

Ale odbočme do histórie našej vedy. Ak odhliadneme od niekoľkých izolovaných jednotlivcov (hocijako i vynikajúcich), systematicky sa naša vedecko-výskumná základňa začala budovať až po oslobodení. Vtedy sme začínali prakticky bez skúseností, so skromnou literatúrou, takmer žiadnym kontaktom so svetom a s veľmi malým počtom osobností, od ktorých by sa bolo možné naučiť vedeckému remeslu. Áno, remeslu. Veď veda, to nie sú len vedomosti, naučené z kníh, ale aj „ako na to", čo sa najlepšie učí systémom majster-učeň, vo vede kodifikovaným v podobe školiteľ-ašpirant. Prečo by inak existovali vedecké školy? Zriedkavé boli a sú prípady, aby sa vedec vypracoval na špičkovú úroveň celkom sám, izolovane. Vzhľadom na nedostatok skúseností, učiteľov a so skromnými možnosťami ísť na skusy do sveta, asi vtedy nebola iná možnosť, ako robiť spočiatku aj výskum tréningový, sledovateľský, teda taký, ktorý sa zaoberá vo svete už vyriešeným problémom. Bolo treba dobehnúť svet, naučiť sa techniky, ktoré už inde vedia, a teda skúmať aj to, čo inde už vyskúmali.

Azda by sme toto štádium aj boli postupne prekonali. Každá ďalšia generácia vedeckého dorastu sa mohla podučiť od predchádzajúcej a začať o stupienok vyššie. Lenže my sme si zaumienili urobiť to akosi prirýchlo, horúcou ihlou. Naše vedecké inštitúcie – vých chatiek na tisíc obyvateľov všetko, čím sa meria vyspelosť národa? Nie je našou povinnosťou dať svetu v kultúre, vede, technike to, čo našej úrovni zodpovedá? Utvoriť našim talentom podmienky na to, aby svetu dali to, čo mu dať môžu?

Nejde iba o nejakú abstraktnú morálnu povinnosť. Dlhodobe a na úrovni môže totiž zo svetovej vedy čerpať iba ten, kto do nej prispeje svojím dielom. Kto sa na tvorbe vedeckých výsledkov nezúčastňuje, zaostane. Veď vedecké informácie sa šíria nielen tlačeným slovom, ale predovšetkým osobným stykom a spoluprácou. Pasívnou účasťou na prednáškach, či konferenciách sa možno málo dozvedieť. Treba sa pýtať, konfrontovať stanoviská, prispieť myšlienkou do mlyna. Myslíte, že niekoho baví odpovedať na nezasvätené otázky? Aj zahraničná stáž je prakticky bezcenná, ak stážista nie je na takej úrovni, aby bol platným členom tohto výskumného kolektívu. S tým súvisí aj otázka zahraničných ašpirantúr, o ktorých hovoril prof. Pišút (NS č. 29). Zriedka sa stáva, že by sa niektorý náš ašpirant dostal k špičkovému školiteľovi. Prečo by aj. Prečo by si významnejší školiteľ mal vziať za ašpiranta mladíka, o ktorom nič nevie, keď si z domácich môže vybrať tých najlepších z najlepších? Iné však je, keď mu renomovaný odborník niekoho odporúča. Pravda, musí to urobiť zodpovedne.

Ako to všetko súvisí s doktorátmi vied? Aj keď nemožno povedať, že titul doktora vied sploštenému chápaniu vedy celkom podľahol, predsa len je ním poznačený. Predstavte si pracovníka, ktorý roky na vedeckom mieste pracuje, robí to, čo sa od neho žiada a čo sa označuje za vedu. Nezačne postupne veriť tomu, že dosiahol vedecký výsledok? Prečo by mal brať za bernú mincu práve text vyhlášky o vedeckých hodnostiach, ktorá celkom jednoznačne tvrdí, že doktorát má značiť prínos do svetovej pokladnice vedy?

Na obranu doktorátov vied však možno povedať, že zo všetkých vedeckých a vedecko-pedagogických titulov sú azda najmenej postihnuté. Nebolo by škoda nechať ich ďalej devalvovať? Pritom je zaujímavé, že vecne vlastne problém ako rozoznať vedeckú osobnosť ani problémom nie je. Vedecké obce jednotlivých vied zasa nie sú také veľké, aby sa v nich nevedelo, kto naň má, kto sa stal doktorom vied za svoje vedecké dielo a kto z iných príčin. Problém je skôr technického rázu: ako vyjadriť, že niekto je vedeckou osobnosťou pomocou formálnych, vyčísliteľných kritérií. Kvôli každému doktorátu predsa nie je možné konať referendum.

Aká je však prax? Naše obhajoby sú väčšinou

prirýchlo, horúcou ihlou. Naše vedecké inštitúcie – akadémia, vysoké školy, výskumné ústavy – rástli a rástli. Zabudli sme, že päťmiliónové Slovensko nemôže vyplodiť neobmedzené množstvo vedeckých talentov a možnosti jedného vedca, akokoľvek vynikajúceho, vyškoliť ďalších tiež nie sú neobmedzené.

Hovoríme, že veda je dnes výrobnou silou, ale chápeme to akosi naruby. Ako keby veda mala výrobu suplovať. Kladieme bezmyšlienkovite rovná sa medzi vedu a výskum – dokonca aj vývoj. Tak ako klavirista nemusí byť skladateľom (hoci obidvoch si treba rovnako vážiť), ak niekto vo svojej výskumnej, či vývojovej práci uplatňuje vedecké poznatky, neznačí to ešte, že je vedcom. Napríklad, ak sa na našej súčiastkovej báze vyvinie počítač s dobrým programovým vybavením, ktorého sériový výrobok je dostatočne spoľahlivý a uspokojuje užívateľov, je to pre spoločnosť cenné a zasluhuje si primerané ohodnotenie. A to aj vtedy, ak počítač nie je porovnateľný so svetovou špičkou. Ale za vedecký výsledok ho možno kvalifikovať iba vtedy, ak sa pri jeho vývoji použili nové, vo svete neznáme princípy. Stručne povedané: realizačný výsledok môže byť pre spoločnosť v danej fáze cennejší ako výsledok vedecký a vyžaduje si primerané ohodnotenie. Ale nie ako výsledok vedecký! Napokon, v platovej a kvalifikačnej stupnici pracovníkov výskumu sa tento rozdiel zreteľne odrazil zavedením vedeckotechnických kvalifikačných stupňov: vedúci vedeckotechnický pracovník môže bez vedeckej hodnosti dosiahnuť plat samostatného vedeckého pracovníka.

Už vidím vztýčený prst, že chcem odtrhnúť vedu od praxe. Vôbec nie. Ide len o to, aby vedci pomáhali praxi vedeckou a nie inou činnosťou. Ani nechcem povedať, že sa vedecké pracoviská nemajú starať o realizačný a sledovateľský výskum. Veď od väčšiny tisícov pracovníkov našej vedecko-výskumnej základne ani reálne viac čakať nemožno. Ale je rozumné vedu takto redukovať? Ak aj tým pracovníkom, ktorí sú chopní na viac, zaplníme všetok čas tým, že im do poslednej čiarky naplánujeme, čo majú vyvinúť (lebo aj Američania, či Japonci to už urobili, alebo to majú v pláne), a navyše im to každé tri roky zmeníme, odpílime im vlastne konár, ktorým sú pre spoločnosť užitoční: prísť s čímsi novým, neočakávaným, čo vyžaduje roky práce a čo je nové nielen pre nás, ale je zároveň trvalým vkladom do svetovej pokladnice vedy.

Ako uvádza prof. Špaňár (NS č. 26), objavujú sa i názory, že sme malý národ a o také méty sa ani snažiť nemôžeme, a ani si to nemôžeme dovoliť. Čo by sme už my tak mohli dať svetovej vede! A načo by sme to vlastne robili – veď z nej stačí čerpať... Lenže my už nie sme ten malý národ, utláčaný a chudobný. Máme mohutný priemysel, vyspelé poľnohospodárstvo a vysokú životnú úroveň. Je však vybavenosť domácností farebnými televízormi, či počet víkendo-

Aká je však prax? Naše obhajoby sú väčšinou nudné. Ako by už o doktoráte bolo rozhodnuté tým, že sa uchádzač k obhajobe pripustí. Obhajoba vyzerá skôr ako oslava uchádzača. Kritický posudok je výnimkou a skôr sa chápe ako keby si oponent s uchádzačom vybavoval osobné účty. Poznám niekoľko prípadov, keď predseda komisie usúdil, že práca nemá úroveň doktorskej práce. Namiesto toho, aby to priamo povedal uchádzačovi (na čo mu vyhláška dáva právo, tak ako dáva právo uchádzačovi trvať na ďalšom pokračovaní), označil prácu za nepatriacu do danej komisie. Nasledoval nechutný „ping-pong" s prácou medzi rôznymi komisiami, ktorý sa skončil vymenovaním komisie ad hoc, hoci vecne na to ani neboli dôvody... Je takéto konanie dôstojné predsedu komisie?

Ale neviňme celkom komisie. Aj ich členovia sú ľudia zraniteľní, podliehajúci tlakom robiť úľavy, výnimky. Čo nám treba? Zmena atmosféry, aby si členovia komisie spoločensky nemohli dovoliť prepustiť slabého uchádzača a aby sa oponenti hanbili na prácu nízkej úrovne napísať priaznivý posudok. Táto atmosféra sa zo dňa na deň zmeniť nedá. Preto sa ešte určitý čas bez vyčísliteľných, objektívne preukázateľných kritérií nemôžeme obísť. Z týchto kritérií je kritérium citácií, ktoré navrhuje prof. Ebringer, podľa môjho názoru, najlepšie. Treba však pritom mať na pamäti, že veda a vedecké výkony sú niečím príliš mnohostranným a subtílnym na to, aby ich bolo možné merať číselnými ukazovateľmi. Inak by sa totiž dali pozbierať dáta, vložiť do počítača, ktorý by vydal verdikt. Ale nemá uchádzač o vedeckú hodnosť preukázať aj schopnosť obhájiť svoje výsledky a tézy v ohni kritiky a konfrontácie s inými názormi?

Doc. Š. Jánoš sa vo svojom diskusnom príspevku zamýšľa nad tým, či vôbec má zmysel trápiť sa kvôli doktorátom – veď v mnohých vedecky vyspelých krajinách sa špičková veda robí aj bez podobnej vedeckej hodnosti. Ja si myslím, že škoda by ich bolo vzdať sa. Doktoráty majú totiž oproti iným vedeckým a vedeckopedagogickým hodnostiam jeden klad: pripúšťajú prístup verejnosti do pokračovania pri ich udeľovaní. Ale túto možnosť dostatočne nevyužívame. Je to naša chyba.

Neviem posúdiť situáciu v iných vedách, ale napríklad v matematike situácia zďaleka nie je beznádejná. Aj tu sa urobili chyby, ale nie je ich ešte tak veľa, že by sa úroveň doktorátov nedala udržať na solídnej úrovni, ba dokonca aj zvýšiť. Je to v silách nás samotných za dvoch predpokladov: ak nebudeme robiť nijaké ďalšie výnimky a naše zodpovedné orgány nás v tejto snahe podržia.

Položme si teda latku vyššie a predovšetkým – prestaňme ju konečne podliezať. Je na to najvyšší čas.

RNDr. PAVOL BRUNOVSKÝ, DrSc.
Ústav aplikovanej matematiky
a výpočtovej techniky UK

P. Brunovský

Kam kráčaš academia

Národná obroda, 20.10.1990.

Kam kráčaš, academia?

Zložité cesty k „academic excellence"

Tentoraz nepôjde o to, kam s Akadémiou, preto aj malé „a" v slove academia. Vzťahuje sa totiž na akadémiu v pôvodnom zmysle slova, teda na obec učených bez ohľadu na to, či pracujú na Akadémii vied, vysokej škole alebo inde.

Demokratizácia neobišla ani akademickú obec. Ako jej odnože vznikli akademické fóra, ktoré prerástli v akademické senáty i Radu vedcov. Keďže fungujú už nejaký čas, nezaškodí zamyslieť sa nad tým, čo priniesli.

V demokracii dávajú ľudia svoj hlas tomu, či onomu. Bolo by naivné myslieť si, že pritom nehľadia na svoje osobné záujmy. Tie sa kryjú s celospoločenskými záujmami natoľko, nakoľko je široký obzor hlasujúcich. Demokracia vychádza z toho, že jestvuje korelácia medzi záujmami väčšiny jednotlivcov a dobrom spoločnosti ako celku.

Vráťme sa však k akademickej obci. Tu je situácia trocha iná, pretože jej členovia prechádzajú výberom. Netreba azda nikoho zvlášť presviedčať, že prvoradým kritériom výberu po dlhé roky neboli odborné predpoklady k vedeckej, či pedagogickej práci. Možno teda od nich očakávať, že nimi demokraticky prijaté uznesenia budú v súlade so spoločenským záujmom?

Spoločenský záujem je nebezpečný pojem, ktorý bol v minulosti neraz zneužitý. Pokúsme sa ho teda aspoň obrazne definovať.

Daňoví poplatníci vkladajú do štátom podporovaného vyššieho vzdelania a vedy nemalé prostriedky. Za to majú právo od členov akademickej obce žiadať vysokú profesionalitu a snahu po tom, čo sa v angličtine veľmi výstižne nazýva „academic excellence" (mimochodom, nenasvedčuje to niečomu, že v slovenčine zodpovedajúceho pojmu

Voľby na ústavnej úrovni vedú k druhoradému vedeniu. Zvolení vedúci síce nebývajú najhorší, akých si možno predstaviť, určite neprivedú ústav do demoralizujúceho chaosu, ale budú určite kompromisníkmi, ktorí nemôžu ukázať cestu k výšinám (excellence), ani ich udržať. Na druhej strane, ak ústav degeneroval do uniformnej priemernosti, urobiť kohokoľvek z jeho stredu vedúcim značí zachovať nízku kvalitu. Nie je iný liek, ako angažovať energického človeka zvonku administratívnou intervenciou.

Ako by asi v Halmosovom hodnotení dopadli naše pracoviská? Najlepšie bude pozrieť sa na to, ako vyznievajú ich demokraticky prijímané rozhodnutia.

V prvej fáze boli nesporne prínosom. Veľmoži akademického sveta boli nezriedka natoľko zlí a skompromitovaní, že zbaviť sa ich bolo potrebné ako z pohľadu spoločnosti, tak aj z pohľadu väčšiny akademickej obce. Teraz je situácia iná. Záujmom spoločnosti je pozdvihnúť úroveň akademických pracovísk. Je to v záujme ich zamestnancov?

Neklamné znaky poukazujú na to, že priemer veľmi rýchlo pochopil nebezpečenstvo, ktoré mu hrozí, a sformoval sa do obranného postavenia. Ako môže akademická úroveň pracoviska ležať na srdci pracovníkovi, voliacemu za svojho zástupcu kolegu, ktorý už má penziu na krku a sotva stihol obhájiť dizertáciu prípadne ju ani nemá, alebo študentovi, ktorý má starosti, aby sa prešmykol do vyššieho ročníka?

ani niet?). Podobne ma návštevník ligového futbalového zápasu právo od hráčov vyžadovať, aby predviedli futbalové umenie a aby zo seba vydali všetko. Možno toto očakávať od akademickej obce? Áno – ak jej úroveň dosahuje určitú kritickú hodnotu. Ak nie, pripomína mužstvo, ktorého hráči sa uspokojili s miestom v strede tabuľky a hrajú akurát toľko, aby si zabezpečili dobré a pohodlné životytie. Myslíte, že by hlasovali za trénera, ktorý ich bude prehánať?

Prv, než sa pokúsime rozhodnúť, či naša akademická obec túto úroveň dosahuje, prečítajme si z autobiografie známeho amerického matematika P. R. Halmosa staf, v ktorej sa zamýšľa nad voľbami vedúcich katedier (v pôvodine department chairman).

Ako by mal byť určovaný vedúci? Na mnohých univerzitách, pravdepodobne na väčšine, matematická obec určuje svojho vedúceho voľbami.

Logaritmické pravidlo A. Weila (prvotriedni ľudia volia prvotriednych ľudí, ale druhotriedni volia treťotriednych) platí rovnako, či ide o doplňanie pracoviska novými pracovníkmi, alebo o voľbu kolegu do vedenia. Možno, že Harvard alebo Chicago sú dosť dobré na to, aby neurobili väčšiu chybu, ale o väčšine pracovísk to neplatí. Vedúci môžu byť volení pre popularitu („dobre vychádza s ľuďmi"), alebo pre subtílne sľuby v kampani („nemyslíš, že je nefair oceňovať výskum viac ako výuku pri určovaní výšky platov?"). O profesionalite, múdrosti a zápale ľudia hovoria, ale väčšinou ich nevolia.

Natíska sa teda otázka, či môžeme očakávať, že akademická komunita dokáže vlastnými silami prelomiť začarovaný kruh priemernosti a prekročiť kritickú úroveň, ktorá zabezpečí, že sa samočinne začne dvíhať k „academic excellence".

Administratívne metódy sme zavrhli, lebo s nimi z minulosti máme zlé skúsenosti. Hoci sa situácia zmenila, nebezpečenstvo ich zneužitia dosiaľ nepominulo. Treba sa však zamyslieť nad tým, prečo ich Halmos v amerických podmienkach odporúča.

V Amerike ako daňový poplatník, tak i študent, ktorý si štúdium platí, veľmi ostro sledujú, ako sa jeho doláriky použijú. Ak je univerzita slabá, veľmi rýchlo to pocíti na odlive študentov, a tým aj na svojich finančných príjmoch. Preto má aj administratíva univerzity záujem na tom mať kvalitných vedúcich katedier.

Pre nás z toho plynie poučenie, že nádej našej vedy je v mládeži, v študentoch. Nechceme, aby za štúdium platili. Azda ich však život raz priprie a presvedčí ich o tom, že štúdium, či ašpirantúra nie sú iba predĺžením bezprostrednej mladosti, ale najmä prípravou na profesionálnu dráhu, na ktorej kvalite záleží. Dúfajme, že nie všetci to riešia tak, že jednoducho pôjdu študovať do zahraničia. Dúfajme, že sa zobudia zo spánku, do ktorého ako keby po vypätí nežnej revolúcie upadli, a že sa tohc, čo im patrí, budú domáhať s rovnakou rozhodnosťou a vtipom, ako vo svojom vystúpení minulej jesene.

PAVOL BRUNOVSKÝ

P. Brunovský

Parazitológia vedy

In: O tvořivosti ve vědě, politice a umění II. Brno, Nadace Universitas Masarykiana, 1993, 356 s.

Parazitológia vedy

Ako krásne to vyzeralo koncom roku 1989! Vedeckí pracovníci, pre ktorých veda bola povolaním, sa konečne zbavili diktátu tých, ktorí vďaka svojmu spoločenskému postaveniu žili z ich výsledkov. Nádejali sme sa, že sa podarí vymyslieť systém riadenia vedy, vďaka ktorému pôjdu prostriedky navždy kam majú a ktorý zabráni prižívovať sa na spoločenskej podpore vedy tým, ktorí majú veľké ústa a dokážu sa zašmajchlovať mocným tohoto sveta.

Dnes je už zrejmé, že sme sa mýlili. Tak trocha s hanbou si my, vedci, musíme priznať, že sme pozabudli na prírodné zákony, ktoré rovnako ako pre iné prírodné spoločenstvá platia aj pre spoločenstvo vedcov. Zabudli sme na to, že parazit je niečo prirodzené, že je prítomný prakticky v každom organizme, v každom spoločenstve. A že na jeho potieranie treba venovať určité prostriedky a určitú námahu.

Ak by sme z dopravy chceli celkom vylúčiť nehody, museli by sme ju celkom odstaviť. Potom by ľudia síce nehynuli pri dopravných nehodách, ovela viac by ich však prišlo o život preto, že by sa k nim nedostala potrava. Podobne, nadmerná snaha o čistotu vedy by bola na úkor vedeckej tvorby.

Darmo, musíme si privyknúť na myšlienku, že v spoločenstve existuje prirodzená rovnováha medzi parazitmi a produktívnymi jedincami. Parazit totiž nemôže celkom zničiť jedincov, na ktorých parazituje, lebo by napokon sám zahynul.

Presvedčivejšie by pravdaže bolo, keby som túto všeobecne známu pravdu doložil jednoduchým matematickým modelom. Na prvý pohľad to ani nevyzerá také ťažké, veď pre spoločenstvo hostiteľ - parazit taký model aj existuje a je totožný s Volterrovým - Lotkovým modelom dravec - korisť. V prípade vedy je však na mieste skôr analógia s burinou a zrnom, kde sú spoločenstvá v konkurenčnom vzťahu. Aby sme dospeli ku koexistenčnej rovnováhe, museli by sme do modelu zahrnúť spätnú väzbu závislosti podpory vedy od vedeckej produkcie. A tá je v prípade vedy osobitne ťažko postihnuteľná.

Úvahy nad modelom však neboli celkom zbytočné. Uvedomil som si totiž, že táto spätná väzba má časovú konštantu, ktorá sa

ráta možno až na desaťročia. Preto vplyv počiatočného stavu dlho pretrváva a spoliehať sa na pôsobenie spomínanej spätnej väzby nemožno. Netreba ani osobitne zdôrazňovať, ako je u nás vďaka štyridsaťročnému výberu ľudí do vedy tento počiatočný stav nepriaznivý.

Nie že by paraziti vo vede prestavovali niečo nového. Ibaže voľakedy ťarcha rozhodovania ležala na mecenášoch, ktorí vedu a učenosť podporovali. Podľa toho, akí boli múdri, podporovali naozajstných učencov, alebo šarlatánov. Ich miesto dnes prevzal štát a vedci sa domohli samosprávy. Nie sú teda už natoľko závislí od nálad moci, ale platia za to častou svojej kapacity - musia písať dizertácie a projekty, posudzovať ich atd. Námaha, ktorú na to musia vynaložiť je o to väčšia, o čo je vedecká komunita menšia. Mnoho činností je totiž rovnako časovo náročných bez ohľadu na veľkosť komunity. Rozdelenie už aj tak primalej česko-slovenskej vedeckej komunity nám v tomto ohľade určite nič neuľahčilo.

Nečudo, že tvoriví vedci sa len neochotne zmierujú so stratou drahocenného času na ich vlastnú vedeckú činnosť, ktorú pre nich predstavuje písanie projektov a posudkov, alebo vysedávanie v komisiách. Vítajú každú zámienku, ktorá im umožní vyhnúť sa takejto činnosti. Na to však parazit, ktorému tvorivá práca nechýba, iba čaká. Postupne obsadzuje riadiace miesta a keď dosiahne väčšinu, "demokraticky" znechutí tvorivým pracovníkom život tak, že títo vidiac márnosť svojej činnosti naozaj už majú dobré dôvody sa jej vzdať.

Treba však dodať, že aj keby všetky obhajobné komisie, všetky grantové a vedecké rady pozostávali z tých najpovolanejších, samo o sebe to ešte nič nerieši, pokiaľ sa ich členovia nenaučia rozhodovať a teda občas povedať aj nie. Znižovanie kritérií nie je veľkorysosťou, ak ide o prostriedky spoločenské - veď podporovať mlátenie prázdnej slamy značí odobrať prostriedky tam, kde by priniesli osoh. Rovnako nebezpečné je však vyhýbať sa rozhodovaniu absolutizovaním formálnych číselných kritérií. Na túto tému sa toho už napísalo aj povedalo veľa, tak iba toľko : parazit sa prispôsobí a veľmi rýchlo nachádza spôsob, ako číselné kritériá naplniť formálne, ale nie vecne. Je to ako s hubením buriny - tiež sa nedá spoliehať na to, že prostriedok proti nej bude pôsobiť naveky.

Pred rokmi chodil taký vtip, že komunizmus nevymysleli lekári, lebo by ho boli najprv vyskúšali na myšiach. Vidí sa mi, že pádom komunizmu sme z toho spoločenského laboratoria neunikli, ibaže sme sa stali takými zvláštnymi myšami, ktoré na pokusoch na sebe tvorivo spolupracujú. Bodaj by sa aspoň svet na našich neveľmi úspešných pokusoch poučil! Teraz by si práve mohol všímať, ako "mediocrity" (voľný slovenský preklad "priemer" mi nepripadá dosť výstižný) vie využiť nedostatky demokratických mechanizmov riadenia vedy a neochotu tvorivých vedcov rozhodovať, aby rozpútal ofenzívu za dobytie svojich načas stratených pozícií.

Situáciu mu veľmi uľahčuje aj súčasná rozkolísaná politická scéna a polarizácia spoločnosti. Parazit nemusí strácať čas vedeckou tvorbou, ani sa zaťažovať svedomím a preto je mu ľahko obracať kabát a votrieť sa do priazne tej politckej sily, ktorá je práve na vzostupe. Nerobí mu problém zmeniť sa z normalizátora vedy sedemdesiatych rokov podľa okamžitej potreby na vedca kresťanského, svetoobčianskeho či národniarskeho. Nepohodlného a vedecky úspešnejšieho kolegu je možné zasa zlikvidovať tým, že bol členom tej, či onej strany , alebo s niektorou nich sympatizoval a s inou zase nesympatizoval, alebo aj preto, že sa jeho meno objavilo v Rudej kráve.

Znie Vám to priveľmi pesimisticky? Vedec by nemal prepadať pesimizmu tým, že spozná v prírode zákonitosť, nech by už bola akákoľvek. Ctí ho, ak sa s nou vyrovná a dokáže z nej vyvodiť praktické dôsledky. V tomto prípade záver znie: nedať sa znechutiť tým, že očista vedy od parazitov nepostupuje tak, ako sme sa naivne domnievali a neuvoľniť parazitovi nijakú pozíciu bez odporu. A predovšetkým nebáť sa rozhodovať!

Na záver este jednu poznámku: Ak sa Vám vidí, že ste v článku niekoho (vrátane seba) spoznali, nemýlite sa - ide o podobnosť vôbec nie náhodnú.

Pavol Brunovský

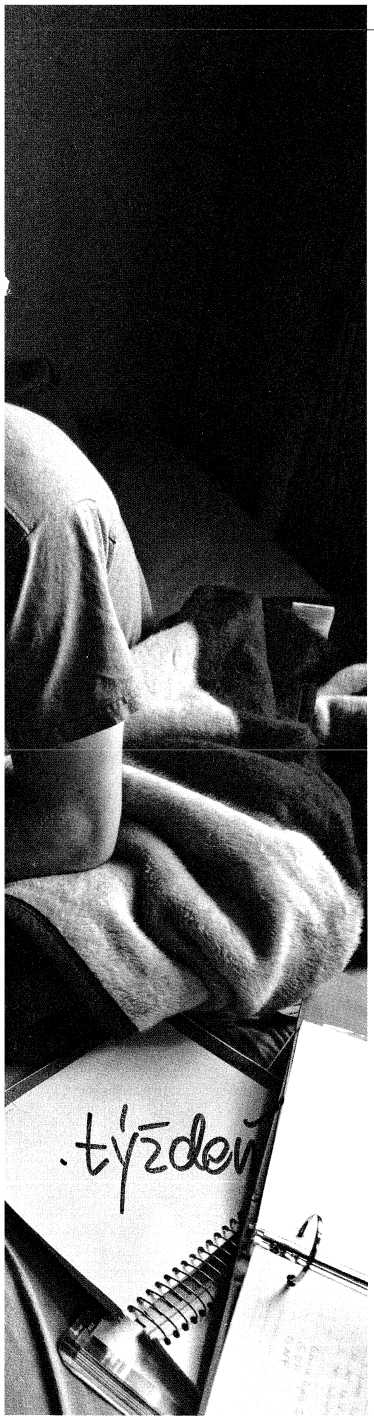
jeseň 1992

Zbornik Tolerovať zlo, vydaný Masarykovou Univerzitou.

P. Brunovský

Koľko stoja reformy?

Týždeň, 14.2.2005.



PAVOL BRUNOVSKÝ

Koľko stoja reformy?

Niekdajšiemu vodcovi Sovietskeho zväzu Nikitovi Chruščovovi stačila letmá návšteva Ameriky, aby sa zhliadol v jej kukurici. A reforma bola na svete: kukuricu treba pestovať od Ukrajiny po Sibír, bez ohľadu na čokoľvek. Permanentný reformný ošiaľ neobchádza ani naše vysoké školstvo.

.Revolučné zmeny priniesli, vďaka veľkorysej autonómii vysokoškolského zákona z roku 1990, zopár rokov relatívneho pokoja. V poslednom čase sme však opäť nabrali tempo.

.kredity ako príklad

Predzvesťou bolo plošné, viac-menej povinné zavedenie kreditového systému. Má príťažlivé myšlienky: dáva študentom možnosť samostatne si tvoriť študijný plán a umožňuje im väčšiu mobilitu. Rýchlo sa však ukázali ohraničenia, ktoré sa dali predpokladať. Jeho plné rozvinutie zastalo na dedičstve komunizmu – priestorovej a organizačnej roztrieštenosti, ktorú si fakulty a katedry žiarlivo strážia. Ako si študent môže zvoliť predmet z ponuky inej fakulty, ak sa vyučuje na druhom konci mesta a on stratí dve hodiny cestovaním? A načo to má robiť, ak sa ten istý predmet vyučuje paralelne na rozličných fakultách, ak nie katedrách?

Predovšetkým sa však ukázalo, že proti zotrvačnej sile masy vysokoškolských pedagógov je akýkoľvek pokus direktívne im vnútiť kreditový systém márny: dnes síce známkujeme stupňami A až Fx namiesto 1 až 4, ale vďaka komplikovanému systému sa spôsob štúdia nijako osobitne nezmenil – možno aj našťastie. Viete si predstaviť, že by si medik nevybral spomedzi voliteľných predmetov anatómiu či strojár mechaniku?

Na prípade kreditového systému možno demonštrovať niekoľko rysov, ktoré sa opakujú pri každom reformnom záchvate. Reformy nevychádzajú zdola, z toho, že by vysokoškolskí pedagógovia, študenti či nebodaj verejnosť pociťovali potrebu zmeny. Namiesto podrobnej analýzy reálnosti uskutočnenia reformy a argumentov pre a proti je tu nejaký odkaz na „autoritu" – či je to sovietske školstvo, entý zjazd strany alebo Bolonská deklarácia. Nikdy sa neskúma, či sa ciele nedajú zabezpečiť lacnejšie a jednoduchšie než za cenu masívnej a administratívne náročnej reformy. Nikdy sa nedajú na misky váh straty z toho, že sa učitelia namiesto vyučovania a bádania venujú písaniu projektov a administratíve, nikdy sa po čase neurobí analýza, čo reforma skutočne priniesla. A predovšetkým, nikdy sa nepočíta so zotrvačnosťou akademickej obce, ktorá si zdanlivo radikálnu reformu prispôsobí tak, aby sa vlastne nič podstatného nezmenilo (čím, mimochodom, nie až tak zriedkavo zredukuje škody, ktoré by dôsledné uplatnenie nepremysleného zásahu prinieslo).

.zrušiť? Nemožné.

Naozajstné tsunami však prišlo až s novým zákonom z roku 2003: povinné zavedenie bakalárskeho stupňa, nová definícia študijných programov a ich plošná akreditácia. Nie že by neboli odbory, kde je bakalársky stupeň odôvodnený. Je však nemálo odborov, ktoré sú prirodzene magisterské, kde vzhľadom na charakter štúdia a tradície nemá bakalársky stupeň opodstatnenie. V dôsledku toho prevažná väčšina bakalárov pokračuje v magisterskom štúdiu na tej istej škole.

Rozčlenenie takého vzdelania na dva stupne je spojené s nemalými nákladmi. Ten istý šťáb pedagógov, ktorý vzhľadom na únik schopných ľudí do iných sfér už dnes zabezpečuje štúdium s odretými ušami, musí navyše vymýšľať a viesť bakalárske práce, zabezpečovať obhajoby a skúšky, väčšinu tých istých študentov opäť nanovo prijímať na magisterské štúdium... A teda opäť bude menej učiť, vzdelávať sa a bádať.

Amerika, podľa ktorej sa bakalársky stupeň v Európe zavádza, nepozná na bakalárskom stupni ani záverečné skúšky, ani záverečné práce. Je celý rad vysokoškolských odborníkov, ktorí na amerických školách učili a vedia to. Dali si tvorcovia zákona námahu opýtať sa ich?

Bakalárskymi programami sa iba zväčšil počet študijných programov, ktorých sú tisíce. A tie má akreditačná komisia posúdiť za dva roky. Na prstoch si možno spočítať, že jednému programu môže venovať iba niekoľko minút. Príprava projektov pre akreditačnú komisiu si vyžiadala obrovskú záťaž pre vysoko kvalifikovaných učiteľov vysokých škôl – veď vypracovanie jedného projektu vyžadovalo niekoľko dní práce.

Načo to všetko, keď dobrá polovica škôl nemá ani personálne, ani materiálne podmienky pre zabezpečenie štúdia na prijateľnej medzinárodnej úrovni? Programy takých škôl by vlastne nemali byť akreditované. Dobre však vieme, že také niečo sa nemôže stať, pretože komisia sa nepresadí proti „politickej vôli" regionálnych, cirkevných či etnických bossov, ktorí si svoju školu nedajú vziať. Veľavravné je čerstvé menovanie troch (!) vysokoškolských profesorov z odboru „technológia vzdelávania". Kto vie vysvetliť, čo to je?

Autor je profesor na Fakulte matematiky, fyziky a informatiky UK v Bratislave

P. Brunovský

... a sme na špici

Týždeň, 2.3.2009.

JAROSLAV DANIŠKA

Nebezpečnejší ako Lisabon?

Írsky podnikateľ a politik Declan Ganley k nám priniesol európsku kampaň. Bolo už na čase – do volieb do Európskeho parlamentu zostáva menej ako sto dní, na uzavretie jednotlivých kandidátok necelý mesiac a na splnenie formalít na zaregistrovanie novej strany asi dva týždne. Ganleyho návšteva však bola užitočná aj z ďalšieho dôvodu. Predstavil sa osobne a predstavil aj projekt celoeurópskej strany. Dojmy sú však zmiešané.

Ganley je charizmatický človek s dobrým prejavom a silným príbehom. Pred rokom sa v Írsku postavil proti všetkým politickým stranám a v úspešnej kampani mobilizoval s podporou cirkvi a tretieho sektora ľudí, ktorí napokon odmietli Lisabonskú zmluvu. Aj podľa .týždňa je to zlá zmluva pre Európu, preto sme jeho úspech ocenili. Malé Írsko ukázalo odvahu: ako jediná krajina, ktorá o upravenej európskej ústave urobila referendum, zmluvu neschválila.

Íri čelili sústredenému tlaku, posmeškom aj arogancii z Paríža či Bruselu, napriek tomu hlasovali slobodne. Tlaku čelia Íri dodnes – referendum si podľa Bruselu aj vlastných politikov musia zopakovať... Je to urážka demokracie a rovnosti Írov s inými Európanmi, ale súčasne príklad, že írski občania majú napriek všetkému väčšie práva ako Francúzi či Holanďania. Paríž aj Amsterdam radšej zmenili vlastné ústavy, len aby nemuseli opakovať referendum, írski politici to neurobili. Ganley má pre to všetko silný mandát: porazil celú politickú triedu. Keď kráča po chodbách Európskeho parlamentu, budí rešpekt. Otázkou preto bolo, ako sa svoju silu chystá využiť.

Mandát však možno aj stratiť. Stačí spomenúť talianskeho profesora a politika Rocca Buttiglione, ktorý sa stal známym, pretože pre svoju vieru nemohol zastávať úrad komisára Európskej únie. Buttiglione bol obeťou nevídanej netolerancie a ako nikto pred ním mohol hovoriť o premene Európy, ktorú v minulosti zakladali kresťanskí politici, ale ktorá dnes práve kresťanov prestáva tolerovať. Taliansky profesor však neskôr zostal potichu – a jeho potenciál sa stratil.

Ganley sa rozhodol inak. Hoci začal ako euroskeptik, ktorému prekáža európska centralizácia a harmonizácia, dnes má reč euronadšenca. Potvrdil to aj v rozhovore, ktorý poskytol .týždňu: neprekáža mu prenos kompetencií do Bruselu, iba nedostatok demokracie. Neprekáža mu, že v Bruseli sa rozhoduje o troch štvrtinách celej legislatívy, ale to, že zákony navrhujú nevolení komisári či ministri, a nie europoslanci. Ľudia, ktorí chcú mať z EÚ superštát, teda robia chybu, že v ňom vidia nepriateľa. On je totiž v skutočnosti prvým politikom superštátu. Verí v európsky ľud, verí v európsku demokraciu, a preto zakladá celoeurópsku stranu. Prekáža mu síce, že Európsky parlament, jediná volená inštitúcia EÚ, je najradikálnejšou zo všetkých inštitúcií v Bruseli. Ale riešenie, ktoré navrhuje, je centralistickejšie ako celá Lisabonská zmluva: chce dať Európskemu parlamentu reálnu moc. Na Slovensku už vieme, čo to znamená. Stačí sa pozrieť, ako sa v Bruseli zmenili naši poslanci – v europarlamente hlasujú za to, čo si doma nikdy nedovolili. Či už je to problematika manželstva, alebo pracovného práva.

Ganley je populista. Je oveľa lepší, keď kritizuje, ako keď niečo navrhuje.

Declan Ganley mnohých, najmä mladých ľudí, na Slovensku zaujal. Sršala z neho energia, nápady a prísľub budúceho „džobu". Ešte jedna vec by však pozornosti ujsť nemala – Ganley je populista. Je oveľa lepší, keď kritizuje, ako keď niečo navrhuje. V negatívnej kampani je šampión, dobrá politika si však vyžaduje viac. Ale keď príde na to viac, Ganley buď odpoveď nemá a dovoláva sa referenda, alebo jeho návrhy zaváňajú revolúciou, a teda neprimeraným rizikom.

Pre strany, ktoré chcú zastupovať v Bruseli oprávnené záujmy vlastnej krajiny, bude lepšie, ak dajú od Ganleyho ruky preč. Nie je to ani Václav Klaus, a už vôbec nie Margaret Thatcher. V Írsku pomohol dobrej veci, viac z neho ale asi nebude. ●

PAVOL BRUNOVSKÝ

...a sme na špici

Dosť už bolo plaču nad tým, ako sú naše vysoké školy na chvoste. Ministerstvo školstva si zrejme povedalo, že plakať nestačí, treba konať.

Už nijaké pokukovanie po všelijakých portugalských modeloch ako v minulosti, treba byť originálny. Piateho júna 2008 vydalo Rozhodnutie. V ňom je záväzne stanovené, aké minimálne podmienky na udeľovanie profesúr musia kritériá školy obsahovať, aby škola dostala akreditáciu. Medzi nimi sú:

a) publikovanie najmenej jednej monografie
b) publikovanie najmenej jednej vysokoškolskej učebnice a dvoch skrípt alebo učebných textov
c) vykonávanie pedagogickej činnosti počas najmenej piatich rokov od získania titulu docent

Odhliadnuc od ďalších podmienok, tie citované stačia na to, že sme sa vyšvihli na čelo. Takéto podmienky totiž nemá nijaká krajina na svete. Profesorom by sa u nás nemohol stať ani Einstein, ani prevažná väčšina ďalších nositeľov Nobelových cien či Fieldsových medailí...

Veci sa medzitým pohli, vedecké rady vysokých škôl kritériá upravujú – niektoré veľmi iniciatívne. Viaceré školy už tieto podmienky do svojich akreditačných spisov zaradili, r. k. bohoslovecká fakulta UK si navyše ešte jednu monografiu pridala. Pravdaže, nájdu sa školy, ktoré majú psychické bariéry. Ako obvykle si svoje vratké sebavedomie upevňujú akýmisi publikačnými a citačnými indexami a pochybnými ratingmi. Na výber však našťastie nemajú, budú sa musieť chtiac-nechtiac prispôsobiť.

V svetle tohto originálneho kroku treba prehodnotiť aj problematiku úniku mozgov. Má nám byť ľúto za Ľubošom Pástorom alebo Petrom Poláčikom, profesormi na University of Chicago či University of Minnesota? U nás by veru podmienky na profesúru nesplnili. Čo už to môže byť za univerzita, ktorá jedného urobí profesorom ako 35-ročné ucho a druhému profesúru priamo ponúkne? Podmienky Rozhodnutia sú našťastie dokonalou hrádzou proti tomu, aby niekto z profesorov týchto takzvaných prestížnych škôl dostal niekedy chuť vrátiť na slovenskú vysokú školu.

Neverím, že by zahraničie tento náš odvážny novátorský krok časom nenasledovalo. Predstavte si tú mohutnú záplavu monografií a učebných textov, ktorá vznikne v 80-miliónovom Nemecku, alebo nebodaj v 300-miliónových Spojených štátoch. Ešte aj papierne ľahšie prekonajú súčasnú hospodársku krízu. A s nimi aj šrotovne.

Autor je matematik ●

.mínus týždňa

P. Brunovský

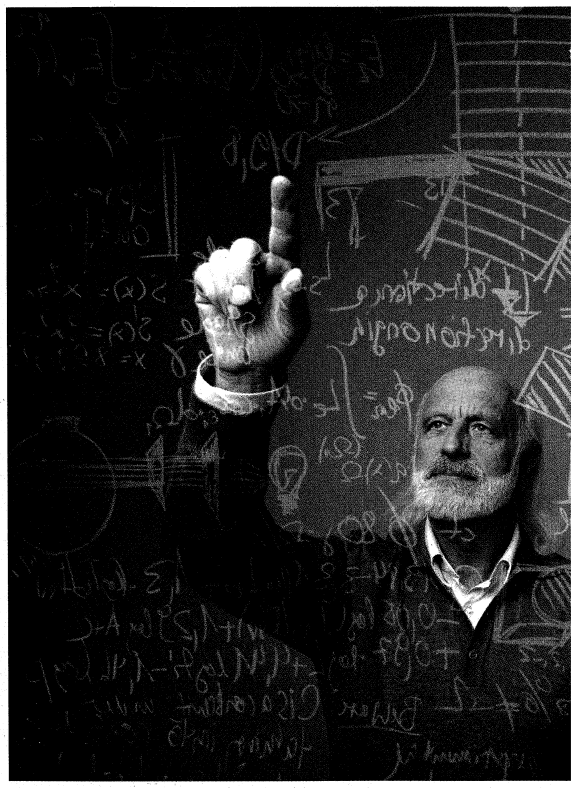
Zasa raz profesori v médiách

Týždeň, 4.7.2011.

PAVEL BRUNOVSKÝ

Zasa raz profesori v médiách

Tentoraz to bolo vďaka ministrovi školstva, ktorému sa (azda právom) nepozdáva úroveň niekoľkých návrhov na menovanie profesorov, a tak má zábrany posunúť ich prezidentovi. Pokrokom je, že sa snaží neporušiť pritom zákon a neriešiť to plošnými prikázaniami, ktoré vždy postihnú viac tých poriadnych. Ale nič podstatné tým nerieši.



O chvíľu tu totiž bude nová kauza a po nej ďalšie. Verejnosť sa opäť nebude vyznať v tom, kto je dobrý a kto zlý. A vnútri akademickej komunity sa namiesto dôležitejších problémov budú opakovane riešiť neriešiteľné problémy, či má profesor mať 50, alebo 300 citácií a či citácie sú to pravé orechové, alebo nie.

A opäť sa nič nevyrieši, lebo riešenie je celkom inde, len ho treba nahlas vysloviť. Ide o to, že profesúra ako prezidentom menovaný štátny titul je prežitok z doby barónov a dvorných radcov. Modernou progresívnou koncepciou je profesor ako funkcia, do ktorej si škola na základe konkurzu podľa vlastných potrieb a finančných možností vyberie najlepšieho z dostupných kandidátov. Na tejto jednoduchej pravde nič nemení ani skutočnosť, že aj v mnohých vyspelých krajinách prežívajú rozličné bizarné spôsoby udeľovania profesúr.

Mimochodom, o čo vyššie renomé predstavuje naša prezidentom menovaná celoštátna profesúra oproti funkcii profesora na MIT, ktorá formálne nijakú platnosť mimo školy nemá? Navyše, princíp rovnakého metra na profesorov rozličných odborov a rozličných škôl sa reálne nikdy nedal dodržať. Z môjho členstva vo Vedeckej rade UK si spomínam na prípad ktoréhosi lekárskeho odboru, ktorý nemal profesora a podľa pravidiel mu hrozilo odobratie akreditácie. Lenže ten odbor sa inde ako na UK neštudoval a kde sa teda sa mali pre spoločnosť nevyhnutní lekári vychovávať? Zo známych dôvodov sa k nám okrem pochybných indivíduí (česť výnimkám) zo zahraničia kvalitní akademici nehrnú, a tak nakoniec nebolo iného východiska, ako zavrieť obe oči a niekoho do tej hodnosti už voľajako preštrikovať.

Skúsme si predstaviť, čo by sa stalo, keby sa profesúra zmenila z titulu na funkciu. Odpadli by ťažkosti s neriešiteľnou snahou o rovnaký meter na profesorov rozličných odborov a rozličných škôl. Minister by nemal pravidelný hlavybôľ z návrhov, v ktorých školy podliezajú svoje vlastné, aj tak nízke kritériá. Vytvoril by sa priestor na iniciatívu škôl sprísňovať nároky na svojich profesorov a tým si zvýšiť renomé. Školy by si samy regulovali počet profesorov tak, aby si ich vedeli zaplatiť. Udeľovanie akreditácie by sa nemohlo viazať na formálne pravidlo prítomnosti menovaného profesora. Tým by zmizol problém lietajúcich profesorov a profesorov zo zahraničia, ktorých študenti uvidia tak raz za semester vo výťahu.

Každá zmena, samozrejme, so sebou prináša isté riziká a potrebu nových pravidiel. Konkurzný systém a autonómia vysokých škôl v obsadzovaní profesorských miest vyžaduje ako samozrejmosť, aby financovanie vysokých škôl bolo celkom nezávislé od počtu ich profesorov. Ale aj tak štát, samozrejme, nemôže možnosť udeľovania profesúr nechať až tak celkom bez dohľadu. Úplná autonómia, ako napríklad v USA, je u nás problematická vzhľadom na to, že školy čerpajú prostriedky výhradne zo štátneho rozpočtu. Garde by sa však obrátilo: akreditácia by nestála na tom, či máš, alebo nemáš profesora, ale akýchže si to ľudí za profesorov vymenoval.

Profesúra ako prezidentom menovaný titul je prežitok z doby barónov a dvorných radcov.

Netreba si zasa robiť ilúzie, že táto zmena je čarovným prútikom. Mnoho problémov je daných prostredím Slovenska, jeho veľkosťou, tradíciami a postojmi verejnosti. Pre rečové a finančné bariéry nemôžeme očakávať, že sa nám na vypísaný konkurz prihlásia desiatky záujemcov z celého sveta ako hoci aj v susednom Rakúsku.

Citlivou otázkou je, či profesorské funkcie majú byť systematizované, teda či ich počet má mať škola určená. To by v podmienkach malého Slovenska, kde sa nezriedka jeden odbor vyučuje na jednej-dvoch školách, bolo veľkým rizikom. Vysoké školstvo je totiž u nás ďaleko od ustáleného stavu, a tak vzniká nebezpečenstvo toho, že obsadenie systemizovaného miesta uchádzačom z núdze by na dlhý čas zablokovalo možnosť angažovania podstatne kvalitnejšieho odborníka, ktorý sa medzičasom objaví. To by len podporilo odliv do zahraničia ešte aj tých, ktorí by si za miesto svojho pôsobenia radi zvolili Slovensko.

Celá vec však má háčik. Úloha prezidenta menovať profesorov je zakotvená v Ústave. Šikovní právnici však už s Ústavou vedeli zariadiť onakvejšie veci.

Autor je matematik ●

P. Brunovský

Zakliaty Kopec

Týždeň, 14.3.2013.

.spoločnosť

Zakliaty Kopec

.pavel Brunovský

Pri pohľade z diaľky môže Kopec pripomínať univezitný campus. Lenže keď od zastávky Botanická záhrada v bratislavskej Karlovej Vsi kráčate do mierneho briežku, tento pocit vás v Mlynskej doline opustí.

zrazu sa vám otvoria obrazy chrastavých budov s padajúcimi balkónmi, ošarpané unimobunky a baraky so záhadnými užívateľmi, hrdzavé ploty z vlnitého plechu, za nimi náletové lesíky a krovie posiate odpadom, v ktorých má poriadok iba bezdomovec okolo svojho stanu.

.jama

Ale ak vytrváte a prejdete kúsok za Kopernika na plochom vrchole kopca, čaká vás pointa: namiesto obrovskej jamy po nedokončenom školskom jadrovom reaktore, ktorá na ňom ešte pred niekoľkými rokmi zívala, stojí nová-novučičká veľkorysá budova Fakulty informatiky a infomatických technológií Slovenskej technickej univerzity s ešte veľkorysejším prázdnym parkoviskom.

Jama zohrala v našej histórii svoju úlohu. Dodnes sa síce vedú spory o tom, či by reaktor nozaj bol zlom, alebo prínosom, ale práve demonštráciou proti nemu sa začal proces, ktorý študentov sformoval ako jednu z kľúčových zložiek Novembra. Ešte donedávna to pripomínal nápis na plote „Keď sa takto býva, reaktor nechýba“.

Niežeby bolo za čím ľutovať. Naopak, bol by to dôvod potešiť sa – keby budova a parkovisko nezívali prázdnotou. Prečo to tak je, o tom sa ku mne dostali iba chýry. Ak by ste si chceli kontrast vychutnať, odporúčam navštíviť Kopec v strede týždňa okolo poludnia. Vtedy je všade – okrem novej budovy – hŕba študentov, a nie je pomaly kde zaparkovať. Betónové zábrany však vstupu na nové parkovisko bránia.

.ako sa to začalo

Keď sa po skončení vojny začali slovenské vysoké školy systematickejšie rozširovať, nebola azda ani jedna z nich v budove, ktorá by bola na tento účel stavaná. Napríklad budova na Šafárikovom námestí, v ktorej bola väčšina Univerzity Komenského, sa pôvodne stavala ako burza, celá Technická univerzita na Vazovovej a Mýtnej zasa sídlila v budovách tuším učňovskej školy.

Povráva sa, že už vtedy sa objavila myšlienka vybudovať na Kopci campus. A že komunistická vrchnosť potom tú myšlienku zavrhla pre strach z výbušnej koncentrácie intelektu. Či je to pravda, alebo nie, prišla ľadová doba a plánom bol na dlhý čas koniec. Na celé desaťročie od roku 1953 sa Kopec dostal do bezprostrednej blízkosti zakázaného hraničného pásma, ktoré vtedy zahŕňalo celú Karlovu Ves. A tak bolo na Kopci naďalej ticho, ktoré namiesto študentov rušili iba škorce, poletujúce nad odvekými vinohradmi.

V tom čase sa začali stavať pre budovy vysoké školy, ale sprvu najmä pre techniky. Situácia sa zmenila až vtedy, keď budovy Prírodovedeckej fakulty museli ustúpiť brutálnemu rezu mestom v podobe rušnej Staromestskej ulice. A tak uzrel svetlo sveta projekt výstavby Prírodovedeckej fakulty UK na Kopci. Nie je prekvapením, že ho zverili architektovi, ktorého klenoty od premostenia galérie cez Archív a Inchebu poznačili Bratislavu na dlhý čas. A ako bolo onému architektovi vlastné, bol to projekt veľkorysý a po dnešných skúsenostiach by sa dalo povedať, že v mnohých smeroch megalomanský.

.stratiť hlas či zadusiť sa?

Je to akurát 40 rokov, čo sa do prvej budovy na Kopci, dnešného Matematického pavilónu „Matfyzu“, nasťahovali matematici a fyzici, ktorí v tom čase ešte patrili do Prírodovedeckej fakulty. Po viac než roku, ktorý strávili pri 6-stupňových teplotách v starej nevykurovanej budove a po útekoch do hostinnejších, ale na vyučovanie nie veľmi vhodných pôsobísk, sa im uľavilo. Ale skoro sa ukázalo, že ich radosť bola predčasná. Veľké posluchárne boli naprojektované bez okien a klimatizácia kompenzovala neúčinnosť hlukom. A tak mal prednášateľ voľbu medzi stratou hlasu z prekrikovania klimatizácie a nedostatkom kyslíka, spojeným s chladom či horúčavou, podľa ročného obdobia.

V menších presklených posluchárňach, ktorým ľudová tvorivosť dala prezývku „akváriá“, bolo namiesto všeobecne používaných žiariviek naprojektované osvetlenie s desiatkami jeho žiaroviek. Vďaka tomu spotreba elektriny úspešne súťažila so spotrebou plynu, ktorým sa veľkoryso vykuroval vesmír cez škáry v nepoddajných železných okenných rámoch, ktoré vďaka vodivým vlastnostiam tohto kovu spoľahlivo odvádzali z budovy teplo. Škáry medzi miestnosťami zasa zabezpečovali dokonalú vnútrofakultnú informovanosť vrátane čuchovej, čo ocenili najmä obyvatelia kancelárií susediacich s toaletami.

.brodenie blatom

Budovy pomaličky pribúdali, ako druhá prišla na rad budova fyziky. Tempo sa natoľko spomalilo, že aj po jej dokončení sa dlho nebolo možné nasťahovať, lebo sa medzičasom zmenili bezpečnostné predpisy.

Postup neurýchlili ani väzni z neďalekej väznice, ktorí ráno príchodzích sprevádzali. Dámy sa ponáhľali domov za svetla, aby sa vyhli predstaveniam exhibicionistov na potemnelých prístupových cestách. Práce sa o niečo pohli iba pred koncom roku, keď bolo treba na prémie zlepšiť plnenie plánu. Na to boli najlepšie zemné práce, a tak sa študenti i učitelia roky v jesenných pľuštiach brodili blatom.

Po niekoľkých takzvaných „volebných“ obdobiach si mestská časť prestala dokončenie areálu dávať ako „volebný záväzok“. A tak aj projekt zostal nedokončený – centrálna budova, v ktorej mali byť fakultná knižnica alebo jedálne, sa už nepostavil. Preto až donedávna Matfyz aj Prírodovedecká fakulta sídlili v rozpore s predpismi v neskolaudovaných budovách a dodnes nemajú adresu. Píšu si paradoxne Mlynská dolina, hoci sú na kopci a dolina, ktorá im prepožičiava meno, je aspoň kilometer ďaleko. Kolaudácia sa uskutočnila iba nedávno, v čase, keď sú budovy také schátrané, že by už pomaly boli vhodné na demoláciu.

V 80. rokoch k budovám pribudla Fakulta elektroniky a informatiky STU, čím spestrila roztrieštenosť mozaiky vysokých škôl na mape mesta. Je síce o dobrý rôčik

Je to akurát 40 rokov, čo sa do prvej budovy na Kopci nasťahovali matematici a fyzici.

mladšia, ale veľmi to nevidno. Naopak, pretože je novej budove FIIT (Fakulta informatiky a informačných technológií) bližšie, kontrast je ešte vypuklejší. Ale aspoň bola včas skolaudovaná, a tak má adresu aj s číslom.

V poslednom roku sa na Kopci opäť začalo hovoriť o nových budovateľských projektoch. Okrem toho, že má ísť o „vedecko-technický" park, tentoraz z dielne vedenia UK, sú obostreté akýmsi tajomstvom. Skeptici vravia, že ho stihne osud podobný novej budove FIIT: že bude stáť oveľa viac, než sa očakáva a že choseň z neho patrí do oblasti snov. Musím sa priznať, že medzi skeptikov patrím aj ja, a to z jednoduchého dôvodu – čo viem, všetky projekty tohto typu na Slovensku tak skončili. Prečo by to malo tentoraz dopadnúť inak?

.stále iný svet

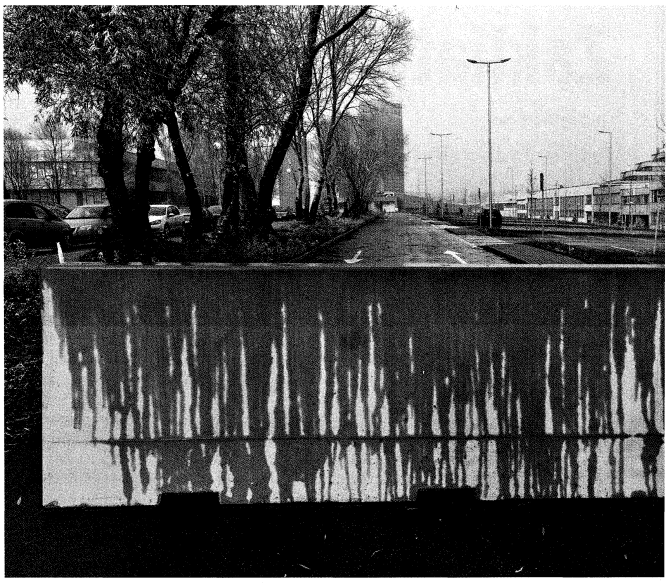
Osobitosť Kopca je však ešte aj v inom. Hoci tomu už dnes málokto verí, fakulty na Kopci fungujú viac-menej tak, ako majú. Normou stále zostáva, že učitelia chodia na prednášky a študentom sa venujú, prijímačky sú čestné a plagiáty sa prísne trestajú. Záverečné práce sú verejne prístupné na internete. Žeby nejaký vplyvný verejný činiteľ za svoje „zásluhy" alebo jednoducho preto, že je „náš", dostal akademickú hodnosť, povedzme, z jadrovej fyziky a podklady sa z archívov stratili, je nemysliteľné. Ak ešte majú naše školy v zahraničí aký-taký cveng, je to v prvom rade zásluha Kopca, na ktorom je dokázateľne prevažujúca akademická sila (čím zasa nechcem povedať, že inde jej niet, len je roztrúsenejšia). Nikde na Slovensku nie je toľko vysokoškolských pedagógov, ktorí majú skúsenosti z dlhodobého profesorského pôsobenia na renomovaných zahraničných pracoviskách vo viacerých kútoch sveta.

Bolo by teda prirodzené, že sa s nimi bude vrchnosť pri prijímaní rozličných opatrení radiť. Boli by sa dozvedeli, že obvyklý argument „v zahraničí je to tak" je zavádzaním. Na nič také si z čias po období ministra Pišúta nespomínam. A to napriek tomu, že spomedzi ministrov a ďalších kľúčových ľudí na ministerstve viacerí z Kopca pochádzali.

Zasa až také prekvapujúce to nie je. Ľudia z Kopca predstavujú pre veľkú časť akademického sveta obrovské nebezpečenstvo, mohli by sa totiž snažiť svoje normy presadzovať aj mimo neho. A tak ich „demokratická väčšina" radšej vytesňuje na okraj a aj jej zástupcovia sa objavujú vo vysokoškolských grémiách veľmi sporadicky a veľmi selektívne. Ak už v nich musia byť z dôvodov pomerného zastúpenia, predstavujú takú malú menšinu, že majú väčšinu „demokratických" hlasovaní vopred prehratú.

Žeby práve to bolo kliatbou Kopca?

Autor je matematik, na Fakulte matematiky, fyziky a informatiky UK založil štúdium ekonomickej a finančnej matematiky. ●



Fotografie v smere hodinových ručičiek: Parkovisko pred Matematickým pavilónom, socha astronóma Kopernika pred „Matfyzom", jedáleň Prírodovedeckej fakulty, Matematický pavilón zvnútra a areál Matematického pavilónu.