Lecture Notes
in Economics and
Mathematical Systems

Managing Editors: M. Beckmann and W. Krelle

287

Dynamical Systems

Proceedings, Sopron, Hungary, 1985

Edited by A. B. Kurzhanski and K. Sigmund

Springer-Verlag

Vol. 184: R. E. Burkard and U. Derigs, Assignment and Matching Problems: Solution Methods with FORTRAN-Programs. VIII, 148 pages. 1980.

Vol. 185: C. C. von Weizsäcker, Barriers to Entry. VI, 220 pages. 1980.

Vol. 186: Ch.-L. Hwang and K. Yoon, Multiple Attribute Decision Making – Methods and Applications. A State-of-the-Art-Survey. XI, 259 pages. 1981.

Vol. 187: W. Hock, K. Schittkowski, Test Examples for Nonlinear Programming Codes. V. 178 pages. 1981.

Vol. 188: D. Bös, Economic Theory of Public Enterprise. VII, 142 pages. 1981.

Vol. 189: A. P. Lüthi, Messung wirtschaftlicher Ungleichheit. IX, 287 pages. 1981.

Vol. 190: J. N. Morse, Organizations: Multiple Agents with Multiple Criteria. Proceedings, 1980. VI, 509 pages. 1981.

Vol. 191: H. R. Sneessens, Theory and Estimation of Macroeconomic Rationing Models. VII, 138 pages. 1981.

Vol. 192: H. J. Bierens: Robust Methods and Asymptotic Theory in Nonlinear Econometrics. IX, 198 pages. 1981.

Vol. 193: J.K. Sengupta, Optimal Decisions under Uncertainty. VII, 156 pages. 1981.

Vol. 194: R. W. Shephard, Cost and Production Functions. XI, 104 pages. 1981.

Vol. 195: H. W. Ursprung, Die elementare Katastrophentheorie. Eine Darstellung aus der Sicht der Ökonomie. VII, 332 pages. 1982.

Vol. 196: M. Nermuth, Information Structures in Economics. VIII, 236 pages. 1982.

Vol. 197: Integer Programming and Related Areas. A Classified Bibliography. 1978 – 1981. Edited by R. von Randow. XIV, 338 pages. 1982.

Vol. 198: P. Zweifel, Ein ökonomisches Modell des Arztverhaltens. XIX, 392 Seiten. 1982.

Vol. 199: Evaluating Mathematical Programming Techniques. Proceedings, 1981. Edited by J.M. Mulvey. XI, 379 pages. 1982.

Vol. 200: The Resource Sector in an Open Economy. Edited by H. Siebert. IX, 161 pages. 1984.

Vol. 201: P. M. C. de Boer, Price Effects in Input-Output-Relations: A Theoretical and Empirical Study for the Netherlands 1949–1967. X, 140 pages. 1982.

Vol. 202: U. Witt, J. Perske, SMS – A Program Package for Simulation and Gaming of Stochastic Market Processes and Learning Behavior. VII, 266 pages. 1982.

Vol. 203: Compilation of Input-Output Tables. Proceedings, 1981. Edited by J. V. Skolka. VII, 307 pages. 1982.

Vol. 204: K.C. Mosler, Entscheidungsregeln bei Risiko: Multivariate stochastische Dominanz. VII, 172 Seiten. 1982.

Vol. 205: R. Ramanathan, Introduction to the Theory of Economic Growth. IX, 347 pages. 1982.

Vol. 206: M.H. Karwan, V. Lotfi, J. Telgen, and S. Zionts, Redundancy in Mathematical Programming. VII, 286 pages. 1983.

Vol. 207: Y. Fujimori, Modern Analysis of Value Theory. X, 165 pages. 1982.

Vol. 208: Econometric Decision Models. Proceedings, 1981. Edited by J. Gruber. VI, 364 pages. 1983.

Vol. 209: Essays and Surveys on Multiple Criteria Decision Making. Proceedings, 1982. Edited by P. Hansen. VII, 441 pages. 1983.

Vol. 210: Technology, Organization and Economic Structure. Edited by R. Sato and M.J. Beckmann. VIII, 195 pages. 1983.

Vol. 211: P. van den Heuvel, The Stability of a Macroeconomic System with Quantity Constraints. VII, 169 pages. 1983.

Vol. 212: R. Sato and T. Nôno, Invariance Principles and the Structure of Technology. V, 94 pages. 1983.

Vol. 213: Aspiration Levels in Bargaining and Economic Decision Making. Proceedings, 1982. Edited by R. Tietz. VIII, 406 pages. 1983.

Vol. 214: M. Faber, H. Niemes und G. Stephan, Entropie, Umweltschutz und Rohstoffverbrauch. IX, 181 Seiten. 1983.

Vol. 215: Semi-Infinite Programming and Applications. Proceedings, 1981. Edited by A.V. Fiacco and K.O. Kortanek. XI, 322 pages. 1983.

Vol. 216: H.H. Müller, Fiscal Policies in a General Equilibrium Model with Persistent Unemployment. VI, 92 pages. 1983.

Vol. 217: Ch. Grootaert, The Relation Between Final Demand and Income Distribution. XIV, 105 pages. 1983.

Vol. 218: P.van Loon, A Dynamic Theory of the Firm: Production, Finance and Investment. VII, 191 pages. 1983.

Vol. 219: E. van Damme, Refinements of the Nash Equilibrium Concept. VI, 151 pages. 1983.

Vol. 220: M. Aoki, Notes on Economic Time Series Analysis: System Theoretic Perspectives. IX, 249 pages. 1983.

Vol. 221: S. Nakamura, An Inter-Industry Translog Model of Prices and Technical Change for the West German Economy. XIV, 290 pages. 1984.

Vol. 222: P. Meier, Energy Systems Analysis for Developing Countries. VI, 344 pages. 1984.

Vol. 223: W. Trockel, Market Demand. VIII, 205 pages. 1984.

Vol. 224: M. Kiy, Ein disaggregiertes Prognosesystem für die Bundesrepublik Deutschland. XVIII, 276 Seiten. 1984.

Vol. 225: T.R. von Ungern-Sternberg, Zur Analyse von Märkten mit unvollständiger Nachfragerinformation. IX, 125 Seiten. 1984

Vol. 226: Selected Topics in Operations Research and Mathematical Economics. Proceedings, 1983. Edited by G. Hammer and D. Pallaschke. IX, 478 pages. 1984.

Vol. 227: Risk and Capital. Proceedings, 1983. Edited by G. Bamberg and K. Spremann. VII, 306 pages. 1984.

Vol. 228: Nonlinear Models of Fluctuating Growth. Proceedings, 1983. Edited by R.M. Goodwin, M. Krüger and A. Vercelli. XVII, 277 pages. 1984.

Vol. 229: Interactive Decision Analysis. Proceedings, 1983. Edited by M. Grauer and A.P. Wierzbicki. VIII, 269 pages. 1984.

Vol. 230: Macro-Economic Planning with Conflicting Goals. Proceedings, 1982. Edited by M. Despontin, P. Nijkamp and J. Spronk. VI, 297 pages. 1984.

Vol. 231: G.F. Newell, The M/M/∞ Service System with Ranked Servers in Heavy Traffic. XI, 126 pages. 1984.

Vol. 232: L. Bauwens, Bayesian Full Information Analysis of Simultaneous Equation Models Using Integration by Monte Carlo. VI, 114 pages. 1984.

Vol. 233: G. Wagenhals, The World Copper Market. XI, 190 pages. 1984.

Vol. 234: B.C. Eaves, A Course in Triangulations for Solving Equations with Deformations. III, 302 pages. 1984.

Vol. 235: Stochastic Models in Reliability Theory. Proceedings, 1984. Edited by S. Osaki and Y. Hatoyama. VII, 212 pages. 1984.

Vol. 236: G. Gandolfo, P.C. Padoan, A Disequilibrium Model of Real and Financial Accumulation in an Open Economy. VI, 172 pages. 1984.

Vol. 237: Misspecification Analysis. Proceedings, 1983. Edited by T.K. Dijkstra. V, 129 pages. 1984.

Lecture Notes in Economics and Mathematical Systems

Managing Editors: M. Beckmann and W. Krelle

287

Dynamical Systems

Proceedings of an IIASA (International Institute for Applied Systems Analysis) Workshop on Mathematics of Dynamic Processes Held at Sopron, Hungary, September 9–13, 1985

Edited by A. B. Kurzhanski and K. Sigmund

Springer-Verlag

Berlin Heidelberg New York London Paris Tokyo

ISBN 3-540-17698-5 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-17698-5 Springer-Verlag New York Berlin Heidelberg

Printed in Germany

Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.
2142/3140-543210

PREFACE

The investigation of special topics in systems dynamics – uncertain dynamic processes, viability theory, nonlinear dynamics in models for biomathematics, inverse problems in control systems theory – has become a major issue at the System and Decision Sciences Research Program of the International Institute for Applied Systems Analysis.

The above topics actually reflect two different perspectives in the investigation of dynamic processes. The first, motivated by control theory, is concerned with the properties of dynamic systems that are stable under variations in the systems' parameters. This allows us to specify classes of dynamic systems for which it is possible to construct and control a whole "tube" of trajectories assigned to a system with uncertain parameters and to resolve some inverse problems of control theory within numerically stable solution schemes.

The second perspective is to investigate generic properties of dynamic systems that are due to nonlinearity (as bifurcations theory, chaotic behavior, stability properties, and related problems in the qualitative theory of differential systems). Special stress is given to the applications of nonlinear dynamic systems theory to biomathematics and ecology.

One of the objectives of the SDS Program was to invite a group of prominent researchers in system dynamics for a workshop that could give an overview of research topics, new results, and open problems in these respective areas of research. Such a workshop on the "Mathematics of Dynamic Processes" was organized by Professor K. Sigmund of the University of Vienna and SDS IIASA in September 1985 at Sopron, Hungary. The proceedings are presented in this volume.

I believe that the workshop has also achieved one of the goals of IIASA, which is to promote and encourage cooperation between the scientists of East and West.

A.B. Kurzhanski
Chairman
System and Decision Sciences Program
International Institute for Applied
Systems Analysis

CONTENTS

I. DISCRETE DYNAMICAL SYSTEMS

Resonant Elimination of a Couple of Invariant Closed Curves in the Neighborhood of a Degenerate Hopf Bifurcation of Diffeomorphisms of $\mathbf{R}^2$

A. Chenciner
Département de Mathématiques, Université Paris VII, 2 Place Jussieu, 75251 Paris Cedex 05, France

By a "standard family of elimination" we mean a one-parameter family N_μ of local diffeomorphisms of $(\mathbf{R}^2,0)$ fixing 0 and invariant under rotation, whose unique bifurcation value μ_Γ corresponds to the coalescence and then disappearance of two normally hyperbolic invariant circles of N_μ (*Figure 1*).

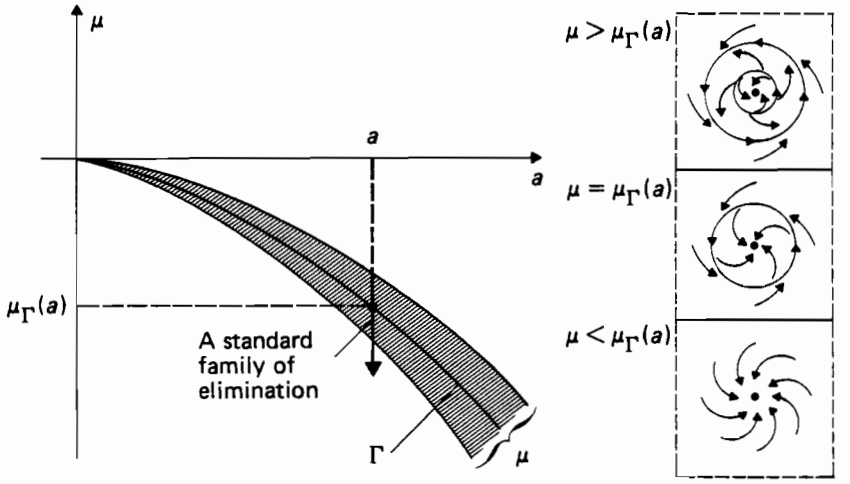


Figure 1

In a series of works (Chenciner, 1982, 1985a,b; briefly described in Chenciner, 1985c), I studied the dynamics of generic two-parameter families $P_{\mu,\alpha}$ of local diffeomorphisms having the following property: restricted to a conic neighborhood $\mathcal{V}$ of a certain curve $\Gamma: \mu = \mu_\Gamma(\alpha)$ in the parameter space, the family $\alpha \mapsto (\mu \mapsto P_{\mu,\alpha})$ appears (in well-chosen coordinates) as a perturbation of a one-parameter family, $\alpha \mapsto (\mu \mapsto N_{\mu,\alpha})$, of standard families of elimination that degenerate at $\mu = \alpha = 0$ and are such that the rotation number of the restriction of $N_{\mu_\Gamma(\alpha),\alpha}$ to its unique invariant circle varies monotonically with parameter α (*Figure 1*). The diffeomorphisms $N_{\mu,\alpha}$ are called "normal forms".

I have shown that these two-parameter families are a good dissipative analogue of a generic area preserving local diffeomorphism F of $(\mathbb{R}^2, 0)$ having an elliptic fixed point at 0. In fact, such a diffeomorphism F can always be written (in well-chosen coordinates) as a perturbation of a normal form that leaves invariant each circle centered at the origin, and such a normal form can be thought of as a one-parameter family of circle rotations degenerating at the origin, where the parameter (= radius of the circle) becomes 0, the rotation number varying monotonically with the parameter ("standard twist").

To make precise what follows, let us define "good" rotation numbers [the circle is supposed of length one and the spectrum of $DF(0)$ – or $DP_{0,0}(0)$ – is $\{e^{2\pi i\omega_0}, e^{-2\pi i\omega_0}\}$]. An irrational number ω is "good" if there exist constants $C > 0$, $\beta \geq 0$, such that, for any rational p/q, one has

$$|\omega - \frac{p}{q}| \geq \frac{C\,|\omega - \omega_0|}{|q|^{2+\beta}}$$

and if moreover $|\omega_0 - \omega_0| \leq \varepsilon(C,\beta)$, where the positive function ε depends only on F (resp. on the family $P_{\mu,\alpha}$). A rational number p/q is "good" if there exists a positive constant C such that $q\,|\omega_0 - p/q| < C$ and $|\omega - \omega_0| < \varepsilon(C)$ where the positive function ε depends only on F (resp. on the family $P_{\mu,\alpha}$). To the F-invariant closed curves of class C^∞, given by K.A.M. theory, on which F is C^∞-conjugated to a rotation R_ω of "good" irrational rotation number ω, corresponds a Cantor set of values of α for which the family $\mu \mapsto P_{\mu,\alpha}$ "looks like" a standard family of elimination. Moreover, at the unique bifurcation point $\mu_\Gamma(\alpha)$, the unique invariant closed curve of $P_{\mu_\Gamma(\alpha),\alpha}$ is of class C^∞, and the restriction to it of $P_{\mu_\Gamma(\alpha),\alpha}$ is C^∞-conjugated to the rotation R_ω {see Chenciner, 1985a, Section 2.3; $P_{\mu,\alpha}$ "looks like" $N_{\mu',\alpha'}$ if, in a uniform [independent of (μ,α)] neighborhood of 0 in $\mathbb{R}^2$, the two diffeomorphisms have the same number of invariant closed curves and the same type of decomposition into basins of attraction and repulsion of 0 and the invariant curves; dynamics on the invariant curves are not required to be similar}.

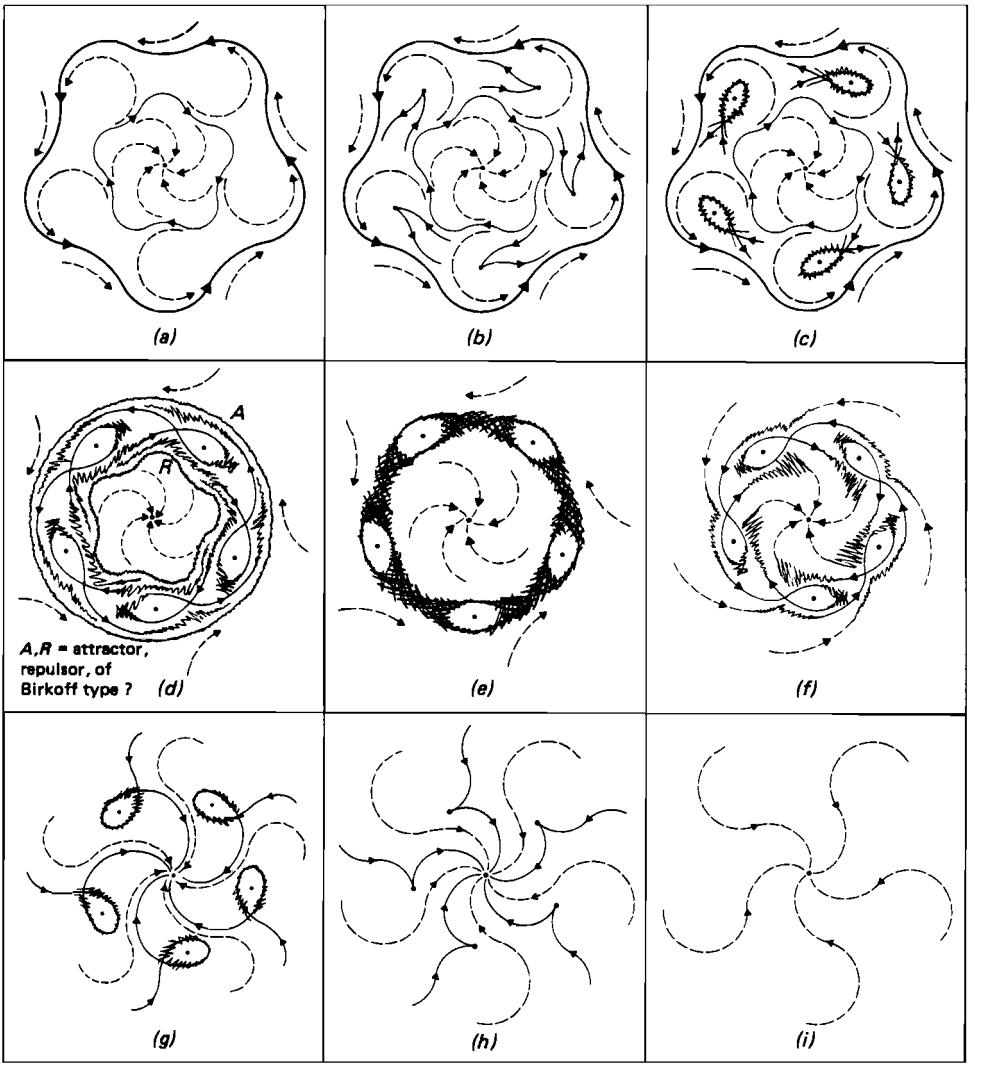


Figure 2 Dynamics of $P_{\mu,a}^{q_n}$ [(*a*), (*b*), (*c*), (*g*), (*h*), and (*i*) are accurate; in (*d*), (*e*), and (*f*) there could exist unexpected invariant curves].

I announce here a similar result for invariant sets whose rotation number belongs to a sequence of "good" rationals p_n/q_n (such a sequence automatically converges to ω_0): to the periodic orbits of F studied in Zehnder (1973), and to their homoclinic orbits, correspond one-parameter subfamilies of the family $P_{\mu,a}$, close to $\mu \mapsto P_{\mu,a_n}$, for which the dynamics look as much as possible like those of a standard family of elimination. The central values of the parameter are associated to a dynamics very close to the one

described by Zehnder (here a_n is such that the restriction of $N_{\mu_1(a_n),a_n}$ to its unique invariant circle is the rotation R_{p_n/q_n}). Such a family is described in *Figure 2* for $q_n = 5$. The remarkable fact is that the only members of the family that do not "look like" members of a standard family of elimination are those that possess a periodic orbit of rotation number p_n/q_n [in the notation of Chenciner, 1985b, those that belong to the Arnold resonance tongue $\hat{C}_{p_n/q_n}$; *Figures 2(b)–(h)*]. In particular, regular invariant closed curves do persist until the appearance inside the ring that they determine of a p_n/q_n-periodic orbit. If the subfamily is well chosen, this orbit will be of Bogdanov type [*Figure 2(b)*]; if not it will be of saddle-node type. The invariant curves persist even further: in *Figures 2(b)* and *(c)* [also *(g)* and *(h)*], the diffeomorphism "looks like" a member of a standard family (a normal form) in the complement of q_n small disks containing the two newborn p_n/q_n-periodic orbits and their homoclinic orbits. Finally, for the central values of the parameter [*Figures 2(d)–(f)*] one can control the invariant manifolds of the hyperbolic p_n/q_n-periodic orbit and show the existence of values of the parameter [*Figure 2(e)*] for which all possible homoclinic intersections occur simultaneously (compare with Zehnder, 1973; in the generic situation a single couple of p_n/q_n-periodic orbits appears).

Figure 3 is taken from Chenciner (1986): it depicts a part of the set of values of (μ,a) such that $P_{\mu,a}$ "looks like" a normal form $N_{\mu',a'}$ (complement of the string of "bubbles", see Chenciner, 1985a, Section 2.3, *Figure 11*): ω_1 and ω_2 are "good" irrationals and the big "bubble" is supposed to intersect the resonance tongue $\hat{C}_{p_n/q_n}$ [set of values of (μ,a) such that $P_{\mu,a}$ possesses at least one p_n/q_n-periodic orbit cyclically ordered as the orbits of the rotation R_{p_n/q_n}; see Chenciner, 1985b]. The subfamily E_n whose dynamics is described in *Figure 2* goes through the region where part of the boundary of this bubble coincides with the boundary of $\hat{C}_{p_n/q_n}$.

Details of the quite long proof appear in Chenciner (1986); the main steps are given here.

1. Obtaining some Periodic Orbits as the Trace of Nearby Resonances

In a generic family $P_{\mu,a}$ it is possible, for a sequence p_n/q_n of "good" rationals converging to ω_0, to follow the bifurcations from the origin of the periodic orbits of rotation number p_n/q_n in the three-parameter family $(\mu,a,t) \mapsto R_t \circ P_{\mu,a}$, t between $(p_n/q_n) - \omega_0$ and 0 (compare with Zehnder, 1973). The p_n/q_n being "good" rationals (which is equivalent to t being small enough) means that $P^{q_n}_{\mu,a}$ is still a perturbation of $N^{q_n}_{\mu,a}$ in an annulus $A_{\mu,a}$ containing the whole nonwandering set of $P_{\mu,a}$ (except 0, of course).

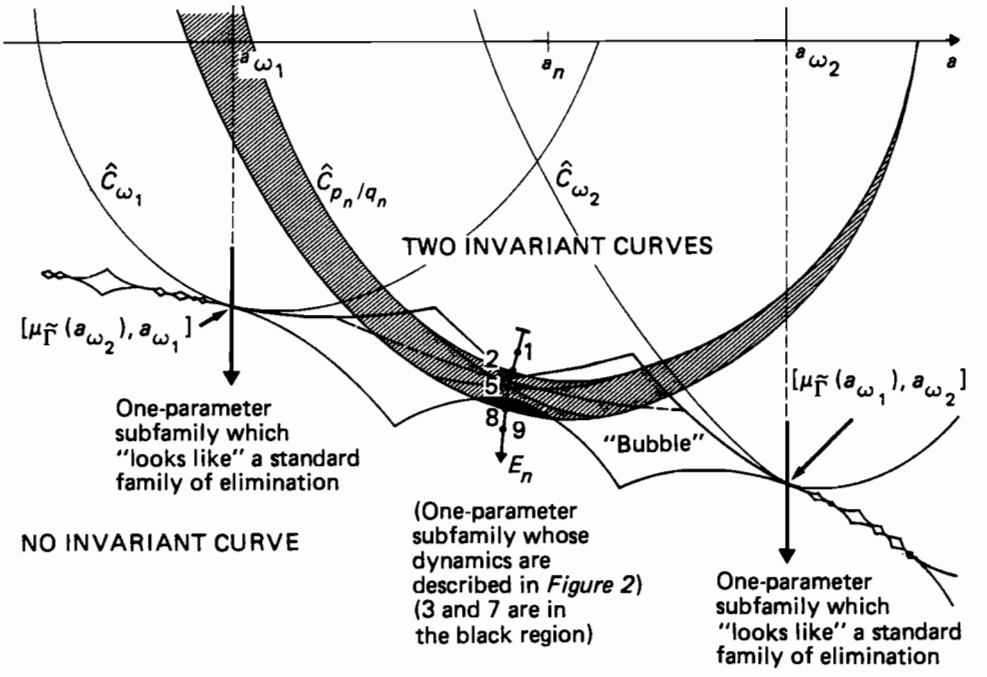


Figure 3 (Compare with *Figure 1*).

2. Putting under Normal Form the p_n/q_n-Periodic Orbits and Approximating the Restriction of $P_{\mu,a}$ to $A_{\mu,a}$ by the Time-One-Map of an R_{p_n/q_n}-Invariant Differential Equation

Let a_n be such that the restriction of $N_{\mu_{\tilde{\Gamma}}(a_n),a_{\tilde{n}}}$ to its unique invariant circle be the rotation R_{p_n/q_n}. When (μ,a) is close to $[\mu_{\tilde{\Gamma}}(a_n),a_n]$, one can choose coordinates (ϑ,y) in the annulus $A_{\mu,a} \approx \mathbf{T}^1 \times [-L,L]$, close to the standard ones, such that:

(1) The restriction of $P_{\mu,a}$ to the set of its p_n/q_n-periodic orbits coincides with the rotation R_{p_n/q_n} (such coordinates always exist if p_n/q_n is a "good" rational but in the generic situation they are easier to obtain because there exist only two such periodic orbits).

(2) $R^{-1}_{p_n/q_n} \circ P_{\mu,a}$ is very well approximated in $A_{\mu,a}$ by the time-one-map of a second-order differential equation $E_{\mu,a}$ invariant under R_{p_n/q_n}:

$$\begin{cases} \dfrac{d\vartheta}{dt} = \omega y \\ \dfrac{dy}{dt} = \alpha + \beta y + \gamma y^2 + \delta\zeta(\vartheta) \end{cases}$$

$\zeta(\vartheta)$ close to $\cos 2\pi q_n \vartheta$ and invariant *under* R_{p_n/q_n}

Moreover, the singularities of $E_{\mu,a}$ are exactly the p_n/q_n-periodic orbits of $P_{\mu,a}$.

3. Showing the Existence of Invariant Closed Curves

Using periodic solutions of $E_{\mu,a}$ and choosing adapted coordinates, one can show by the graph transform method that in one-parameter subfamilies close to $\mu \mapsto P_{\mu,a_n}$, a pair of invariant closed curves exists until (and even after) the birth of a p_n/q_n-periodic orbit (*Figure 2*). The fact that, after this first bifurcation, $P_{\mu,a}$ still looks like a member of a standard family of elimination outside the union of q_n disjoint disks containing the p_n/q_n-periodic orbits that just appeared (and their eventual homoclinic orbits and subordinate tiny closed curves invariant under $P_{\mu,a}^{q_n}$) is proved using Ljapunov functions inspired by the study of the family of differential equations.

4. Controlling Invariant Manifolds of $P_{\mu,a}$ by those of the Hyperbolic Singular Points of $E_{\mu,a}$

This part is in the spirit of Zehnder (1973) with the difference that area preservation is replaced by the existence of the friction parameter β: it has been essentially described by Chenciner (1982, 1983).

5. Conclusion

One salient feature of these results is the analogy between the situations governed by "good" rationals and those governed by "good" irrationals, in particular the persistence as long as this is possible of regular invariant closed curves. In fact, it is the nonexistence of unexpected invariant closed curves in the region of the parameter line where invariant manifolds of the hyperbolic p_n/q_n-periodic orbit present homoclinic tangencies, which looks more difficult to prove (and this even for the family of differential equations). Such a result would show that the process of "resonant elimination" we described involves in its last stage "attractors" reminiscent of those described by Birkhoff (1932).

References

Birkhoff, G. D. (1932), Sur quelques courbes fermées remarquables, *Bulletin de la Société Mathématique de France*, **60**, 1–26.

Chenciner, A. (1982), Points homoclines au voisinage d'une bifurcation de Hopf dégénérée de difféomorphismes de $\mathbb{R}^2$, *C.R.A.S. Série A*, **294**, 269–272.

Chenciner, A. (1983), Bifurcations de difféomorphismes de $\mathbb{R}^2$ au voisinage d'un point fixe elliptique, in Iooss, Helleman, and Stora (Eds), *Proceedings of Ecóle des Houches* (North-Holland, Amsterdam, New York, Oxford).

Chenciner, A. (1985a), Bifurcations de points fixes elliptiques. I: Courbes invariantes, *Publ. Math. de l'I.H.E.S.*, **61**, 67–127.

Chenciner, A. (1985b), Bifurcations de points fixes elliptiques. II: Orbites périodiques et ensembles de Cantor invariants, *Inventiones Math.*, **80**, 81–106.

Chenciner, A. (1985c), Hamiltonian-like phenomena in saddle-node bifurcations of invariant curves for plane diffeomorphisms, in S.N. Pneumaticos (Ed), *Singularities and Dynamical Systems* (North-Holland, Amsterdam, New York, Oxford).

Chenciner, A. (1986), Bifurcations de points fixes elliptiques. III: Orbites périodiques de "petites" périodes et élimination résonnante des couples de courbes invariantes (preprint Université Paris VII).

Zehnder, E. (1973), Homoclinic points near elliptic fixed points, *C.P.A.M.*, **26**, 131–182.

Iterated Holomorphic Maps On The Punctured Plane

J. Kotus
Institute of Mathematics, Technical University of Warsaw, Warsaw, Poland

1. Introduction

The study of the dynamics of complex analytic maps goes back to Fatou and Julia in the 1920s. Their iteration theory of analytic maps was analogous with Poincaré's work. Now this analogy is continued by injecting the modern theory of quasiconformal mappings into holomorphic dynamical systems.

For a holomorphic function f defined in a region (open connected set) U of the closed plane $\bar{\mathbb{C}}$ we denote by f^n the n-fold composition $f^n = f \circ f \circ \cdots \circ f$. A family $\{f^n\}_1^\infty$ is normal if every sequence $\{f^{n_k}\}_1^\infty$ contains a subsequence $\{f^{n'_k}\}_1^\infty$ which converges uniformly on compact subsets of U in the spherical metric. The set of normality for f is

$$N(f) = \{z \in U: \{f^n(z)\}_1^\infty \text{ is normal}\}$$

$J(f) = U - N(f)$ is a Julia set of f. If f is a rational function, then $U = \bar{\mathbb{C}}$. We recall the fundamental properties of a Julia set:

(1) $J(f)$ is a nonempty perfect set.
(2) $J(f^n) = J(f)$ for $n \in N$.
(3) $J(f)$ is completely invariant, i.e., $f^{-1}[J(f)] = J(f)$.
(4) $J(f)$ is the closure of the set of repelling points, i.e.,

$$J(f) = \text{cl}\{z \in U: \exists n\ f^n(z) = z \text{ and } |(f^n)'(z)| > 1\}$$

The properties (1)–(3) were shown by Julia (1918) and Fatou (1919, 1920a,b) for rational maps. In the case of entire transcendental functions, properties (1)–(3) were proved by Fatou (1926), while (4) was proved by Baker (1968). Julia and Fatou investigated the global asymptotic behavior of trajectories of rational maps. Their description was not completed. They did not solve the problem of wandering components of $N(f)$. A component D of $N(f)$ is wandering if $f^n(D) \cap f^m(D) = \phi$ for all $n, m \in N$. Sullivan (1982a,b) solved this problem, proving the nonexistence of wandering components for rational

maps. The theory of iterates of transcendental functions is more complicated. Baker (1976) first showed an example of multiconnected wandering components of $N(f)$ for an entire function. Further, Herman (see Sullivan, 1982a), Baker (1984), and Eremenko and Ljubič (1984a) constructed examples of simply connected wandering components. But Sullivan's theorem is valid for a special class of entire functions S_q (see Nevanlinna, 1970; Gol'dberg and Ostrovskij, 1970): i.e., $f \in S_q$ if there exist $a_1, \ldots, a_q \in \mathbb{C}$ such that $f: \mathbb{C} - f^{-1}(\{a_1, \ldots, a_q\}) \to \mathbb{C} - \{a_1, \ldots, a_q\}$ is a covering map. This result, with additional assumptions, has been obtained by Baker (1984), in a form shown by Eremenko and Ljubič (1984a) and Goldberg and Keen (to appear).

We would like to describe the dynamics for the class R of holomorphic functions $f: \mathbb{C}^* \to \mathbb{C}^*$, $\mathbb{C} = \mathbb{C} - \{0\}$, which has two essential singularities at $\{0, \infty\}$. Since f does not take values 0, ∞, $f(z) = z^k \exp[F(z) + G(1/z)]$, where F and G are nonconstant entire and $k \in N$. Following Baker, we call this class Radström functions, after Radström, who was the first to investigate their properties. Radström (1953) proved properties (1)–(3), but (4) was shown by Battacharyya (1969). Let R_q denote the class of Radström functions f for which there exist points $a_1, \ldots, a_q \in \mathbb{C}^*$ such that $f: C^* - f^{-1}(\{a_1, \ldots, a_q\}) \to C^* - \{a_1, \ldots, a_q\}$ is a covering map. The minimal set of points $a_1, \ldots, a_q$ with this property is called the set of basic points for f. Every basic point a_i is a singularity of f^{-1}. A function $f \in R$ with a finite set of critical values and asymptotic values belongs to R_q for some $g \in N$. The dynamics of $f \in R_q$ is of a mixed type, i.e., admits the properties of dynamics for entire and rational maps. In fact, it admits all types of components described by Sullivan for rational maps and some phenomena which occur for entire maps. Since the fundamental properties of $J(f)$ and $N(f)$ for $f \in R_q$ are the same as in the case of a rational or entire map, the essential question for further investigation of its dynamics is to solve the problem of wandering components. It is expected that in the case of validity of Sullivan's theorem further results concerning the classification of components of $N(f)$, the relation between the number of basic points of f and the number of Fatou domains and Siegel disks, estimation of the Lebesgue measure of $J(f)$, and the problem of J-stability for a generic finite-dimensional family in R_q may be extended. Now we formulate the main results of this paper.

Theorem

If $f \in R_q$, then every component of $N(f)$ is nonwandering.

Proof of this theorem is based on Sullivan's ideas. But it is not immediately clear that this technique may be used for Radström functions. In Section 2 we show how by elimination of different cases the proof of our theorem is eventually based on Sullivan's. In Section 3 we prove other properties of $N(f)$ and $J(f)$ mentioned above for $f \in R_q$. Their proofs are based on results of Eremenko and Ljubič (1984a,b) for entire functions. We describe only new parts. We hope that this is sufficient for readers familiar with iteration theory.

2. Sullivan's Theorem for $f \in R_q$

We begin with some remarks concerning asymptotic values. $\alpha \in \bar{\mathbb{C}}$ is an asymptotic value of $f \in R$ if there exists a path γ tending to 0 or ∞ that satisfies $\lim_{\gamma} f(z) = \alpha$. It is not difficult to prove that, if $f \in R$ and does not take three values in $\bar{\mathbb{C}}$, then f is constant. As a consequence there does not exist any other omitted value for $f \in R$ different from 0 and ∞.

Lemma 1

For $f \in R$ there exist γ_i, $i = 1, \ldots, 4$, such that $\gamma_i \to 0$ for $i = 1,2$, $\gamma_i \to \infty$ for $i = 3,4$, and $\lim_{\gamma_1,\, \gamma_3} f(z) = 0$, $\lim_{\gamma_2,\, \gamma_4} f(z) = \infty$.

Proof

Let $\delta_1 : [0,1) \to \mathbb{C}^*$ be a path tending to ∞, and g be a branch of the inverse function f^{-1}. Then $\gamma_1 : [0,1) \to \mathbb{C}^*$ given by $\gamma_1(t) = g[\delta_1(t)]$ is a path tending either to 0 or ∞. Suppose on the contrary that there exists a sequence $\{\gamma_1(t_n)\}_1^\infty$ that satisfies $\gamma_1(t_n) \to a \neq 0, \infty$ and

$$f(a) = \lim_{n\to\infty} [\gamma_1(t_n)] = \lim_{n\to\infty} f\{[g\,\delta_1(t_n)]\} = \lim_{n\to\infty} \delta_1(t_n) = \infty$$

But f has no pole. If $\gamma_1 \to 0$, then setting $w = 1/z$ in $f(z) = z^k \exp[F(z) + G(1/z)]$ we prove that there exists a path $\gamma_2 \to \infty$ and $\lim h(w) = \infty$ along γ_2 for some other function $h \in R$. To prove the last part of this lemma, it is enough to consider a path $\delta_2 : [0,1) \to \mathbb{C}^*$ tending to 0 instead of ∞.

2.1. Classification of the cases

The following proposition is analogous to Sullivan's. It is trivial in the case of a rational map, but here it is somewhat more delicate. For an entire function $f \in S_q$ it is not necessary to prove it, since in that case any component of $N(f)$ is simply connected.

Proposition 1

If $\{D_n\}_0^\infty$ is an orbit of the wandering domain D_0 of $N(f)$ for $f \in R_q$, then there exists an $n_0 \in N$ such that one of the following cases is satisfied:

(1) D_n has finite topological type and $f : D_n \to D_{n+1}$ is a homeomorphism for $n > n_0$.

(2) D_n is an annulus for $n > n_0$ and there exists a subsequence $\{D_{n_k}\}_1^\infty$

such that the degree of $f^{t_k} : D_{n_k} \to D_{n_{k+1}}$ is greater than 1 for $n_k > n_0$.

(3) D_n has infinite topological type for $n > n_0$.

To prove this proposition we need the following lemma.

Lemma 2

Let $h \in R_p$, D_1 D_2 be components of $N(h)$ with finite topological type. If $h: D_1 \to D_2$ is a covering map and the degree of h is infinite, then each component of $\bar{\mathbb{C}} - D_2$ contains a singularity of f^{-1}.

Proof

Suppose the contrary: then there exists a component of $\bar{\mathbb{C}} - D_2$ which does not contain a singularity of f^{-1}. Let γ be a Jordan curve that lies in D_2 around the boundary of this component and does not include any other component of $\bar{\mathbb{C}} - D_2$. Let B be a simply connected domain bounded by γ. Since f is a covering map and $f^{-1}(z)$ contains infinitely many points, there exist well-defined branches $\{g_k\}_1^\infty$ of the inverse function f^{-1} along curve γ. By the Monodromy Theorem, there exists an extension $\bar{g}_k$ to B that satisfies $\bar{g}_k | \gamma = g_k$. Let $z \in B \cap J(f)$. It follows from the complete invariance of $J(f)$ that $w_k = g_k(z) \in J(f)$. Since $\text{ind}_\gamma z \neq 0$, it follows that $\text{ind}_{g_k(\gamma)} w_k \neq 0$ for any $k \in N$. As a consequence, each domain bounded by $g_k(\gamma)$ contains a boundary point of D_1, i.e., $\bar{\mathbb{C}} - D_1$ contains infinitely many components. This implies that D_1 has infinite topological type.

Proof of Proposition 1

Since f has finitely many basic points, there exists an n_0 such that none of the basic points belongs to D_n for $n > n_0$. We may assume that $n_0 = 0$. Then $f: D_n \to D_{n+1}$ is a covering map for $n \in N$. Suppose that there exists a subsequence $\{D_{n_k}\}_1^\infty$ that belongs to one of the following cases:

(1) Every component D_{n_k} is simply connected. Then D_n is simply connected and $f: D_n \to D_{n+1}$ is a homeomorphism for $n \in N$. Proof of this property is trivial. We leave it to the reader.

(2) Every component D_{n_k} is an annulus. Then D_n is an annulus for $n \in N$. Let D_m be an element of trajectory $\{D_n\}$ that lies between D_{n_k} and $D_{n_{k+1}}$. Then $D_m = f^i(D_{n_k})$, $D_{n_{k+1}} = f^j(D_m)$ for some $i, j \in N$. Since $f | D_n$ is a covering map for $n \in N$, it follows that f_*^i, f_*^j homeomorphisms of fundamental groups induced by f^i, f^j are one-to-one and $z = \pi_1(D_{n_k}) \subsetneq \pi_1(D_m) \subsetneq \pi_1(D_{n_{k+1}}) = Z$. If D_m is a simply connected domain, then $Z \subsetneq \{*\} = \pi_1(D_m)$. Suppose that D_m has finite topological type different from annulus and disk. Then $\pi_1(D_m)$ contains at least the

group $Z \times Z$ and $Z \times Z \subsetneq Z$. These contradictions prove that $\pi_1(D_m) = Z$ for $m \in N$, i.e., D_m is an annulus. Of course, the degree of $f: D_n \to D_{n+1}$ is finite for $n \in N$. Either the degree of $f: D_n \to D_{n+1}$ is 1 for $n > n_0$ or there exists a subsequence $\{D_{n_k}\}_1^\infty$ such that the degree of $f^{i_k}: D_{n_k} \to D_{n_k+1}$ is greater than 1 for $n_k > n_0$.

(3) Every component D_n has finite topological type different from annulus and disk. Then there exist n_0 such that, for $n > n_0$, D_n has the same property and $f: D_n \to D_{n+1}$ is a homeomorphism. Suppose that this is not true: then there exists a sequence of components $\{D_k\}_{k=1}^\infty$, each of which has infinite topological type and D_k lies between D_{n_k} and D_{n_k+1}. Thus the degree of $g_k = f^{i_k}: D_k \to D_{n_k+1}$ is infinite. Otherwise, for covering maps with finite degree d, we have the following rule: $\chi_E(D_k) = d\,\chi_E(D_{n_k+1})$, where χ_E denotes Euler's characteristic. But $\chi_E(D_{n_k+1})$ and d are finite, so $\chi_E(D_k)$ is finite and D_{n_k} has finite topological type. This is a contradiction. As a consequence, the degree of $h_k = f^{j_k}: D_{n_k} \to D_{n_k+1}$ is infinite. By *Lemma 2*, any component of $\overline{\mathbb{C}} - D_{n_k+1}$ contains a singularity of f^{-1}. Since the set of singularities of f^{-1} is finite, infinitely many components D_{n_k} are nested around one singularity of f^{-1}. But D_{n_k} has finite topological type different from disk and annulus, so at least one of the other components of $\overline{\mathbb{C}} - D_{n_k+1}$ contains a singularity of f^{-1}. This implies that only a few components in the sequence $\{D_n\}_1^\infty$ have infinite topological type. We proved also that the degree of $f: D_n \to D_{n+1}$ is finite for $n > n_0$. Otherwise there exists a subsequence $\{D_{n_k}\}_1^\infty$ and maps $\{h_k\}_1^\infty$, $h_k: D_{n_k} \to D_{n_k+1}$, that satisfy the assumptions of *Lemma 2*. But we eliminated this case above. It is not difficult to prove that the degree of $f: D_n \to D_{n+1}$ is 1 for $n > n_1$ using the rule cited for Euler characteristics.

(4) Every component D_{n_k} has infinite topological type. Then D_n has the same property for $n > n_0$. This is a consequence of (3).

2.2. The case of a wandering annulus in $N(f)$ with degree of $f > 1$

Proposition 2

Let $\{D_n\}_0^\infty$ be an orbit of wandering annulus $D_0 \subset N(f)$ for $f \in R_q$. Then the degree of f is equal to 1 for large n.

Proof

Suppose that there exists a sequence of annuli $\{D_{n_k}\}_1^\infty$ such that $f^{i_k}: D_{n_k} \to D_{n_k+1}$ is a covering map and the degree d_k of f^{i_k} is greater than 1. Let ω be a smooth Jordan curve separating boundary components of D_0, and let A_{n_k}, B_{n_k} denote components of $\overline{\mathbb{C}} - D_{n_k}$. Since the family $\{f^n \mid \omega\}_1^\infty$

is normal in the spherical metric, the arc length $f^n(\omega)$ is bounded for $n \in N$. But the degree of $g_k: D_0 \to D_{n_k}$, $g_k = f^{i_1} \circ f^{i_2} \cdots \circ f^{i_k}$, becomes large with k as the product $d_{i_1} d_{i_2} \cdots d_{i_k}$; hence one of the components A_{n_k} or B_{n_k} must have an arbitrarily small spherical diameter for a sufficiently large k. Let A_{n_k} denote this component. Suppose that $A_{n_k} \cap \{0,\infty\} = \phi$ for large k: then $f^{i_k}(A_{n_k}) \cap \{0,\infty\} = \phi$. Since diam $A_{n_{k+1}}$ is small, $B_{n_{k+1}}$ contains 0 or ∞. This implies that $f^{i_k}(A_{n_k}) \not\supset B_{n_{k+1}}$ and, by the Maximum Principle, $f_{i_k}(A_{n_k}) = A_{n_{k+1}}$. Hence A_{n_k} is a sequence of annuli nested around one critical value. But this is impossible. To prove it, it is sufficient to use Sullivan's method (see Sullivan, 1982b). Suppose that $A_{n_k} \cap \{0,\infty\} \neq \phi$ for large k. We assume that $0 \in A_{n_k}$. By *Lemma 1*, there exists a path $\gamma \to 0$ that satisfies $\lim f(z) = \infty$ along γ. On the other hand, for the curve $\omega_k = f^{n_k}(\omega)$, $\text{ind}_{\omega_k} 0$ depends on n_k and tends to infinity for large k. Hence the number of intersections of ω_k with γ becomes unbounded. This implies that the arc length of $f(\omega_k) = f^{n_k+1}(\omega)$ is unbounded for $n_k > n_0$, which contradicts the normality of $\{f^n(\omega)\}_1^\infty$. In the last case (i.e., $\infty \in A_{n_k}$) we use similar arguments.

2.3. The case of a wandering finitely connected component of $N(f)$

We describe in detail the proof that leads to a contradiction in the simply connected case, i.e., if D_n is simply connected and $f: D_n \to D_{n+1}$ is a homomorphism for $n \geq n_0$. The finitely connected, eventually bijective case is virtually identical. Now one has a Riemann map between the region and the sphere minus a finite number of round disks, and their boundary is described by prime ends in the region. The discussion and finally the arc argument are the same.

Proposition 3

There does not exist a wandering component $d_0 \subset N(f)$ for $f \in R_q$ such that D_n is simply connected and $f: D_n \to D_{n+1}$ is a homeomorphism for $n \geq n_0$.

To prove this proposition we need some lemmas. In fact, we show only *Lemma 6* which is new for $f \in R_q$. The others are analogous to the case of entire functions.

In *Lemma 3* we recall some properties of quasiconformal maps: for details see Lehto and Virtanen (1965).

Lemma 3

If $\varphi: G \to H$, $\psi: H \to K$ are quasiconformal, then $\psi \circ \varphi$ is also quasiconformal and the dilatations satisfy:

(1) If ψ is conformal, then $\mu_{\psi\cdot\varphi}(z) = \mu_{\varphi}(z)$ almost everywhere (a.e.).
(2) If φ is conformal, then $\mu_{\psi\cdot\varphi}(z) = \mu_{\psi}[\varphi(z)]\overline{\varphi'(z)}/\varphi'(z)]$ a.e.

If φ,ψ are onto, then we can reverse the implications in (1):

(3) If $\mu_{\psi\cdot\varphi} = \mu_{\varphi}$ a.e. in G, then ψ is conformal.

This implies *Lemma 4*.

Lemma 4

Suppose that f is a one-to-one conformal map from a domain D to a domain D_1, and that φ is a quasiconformal map defined on D and D_1 and whose complex dilatation μ_{φ} is f-invariant, i.e., satisfies $\mu_{\varphi}[f(z)] = \mu(z)f'(z)/\overline{f'(z)}$ a.e. in D. Then $\varphi \circ f \circ \varphi^{-1}$ is conformal in $\varphi(D)$.

Now we formulate the theorem of Ahlfors and Bers (1960).

Lemma 5

Consider a measurable function μ on the plane such that for all t in some open set $T \subset R^n$ one has $\mu(t,z) \in L_{\infty}$ as a function of z with $\| \mu \|_{\infty} < 1$ and that (suppressing z)

$$\mu(t + s) = \mu(t) + \Sigma_{i=1}^{n} a_i(t)s_i + |s|\alpha(t,s)$$

with $\| \alpha(t,s) \|_{\infty} < c$, c constant, and $\alpha(t,s) \to 0$ a.e. in z as $s \to 0$. Suppose that $\| a_i(t+s) \|_{\infty}$ are bounded and that $a_i(t+s) \to a_i(t)$ a.e. for $s \to 0$. Then there exists a unique sense-preserving family of quasiconformal homeomorphisms $\Phi_t = \Phi_{\mu(t)}$ of $\overline{\mathbb{C}}$ onto $\overline{\mathbb{C}}$ such that $\partial\Phi_t/\partial z = \mu(t)\partial\Phi_t/\partial\bar{z}$ a.e. Φ_t fixes 0, 1, ∞ and is in $C^1(T)$ as a function of t for fixed z.

Lemma 6

Let $a_1, \ldots, a_q \in C^*$ be basic points of $f \in R_q$,

$$T = \{(t_1, \ldots, t_{2q+1}) : t_1 \in \mathbf{R}, |t_i| < 1, i = 1, \ldots, 2q+1\}$$

Suppose that Φ_t, $t \in T$, is a family of quasiconformal homeomorphisms of $\overline{\mathbb{C}}$ which fix 0, 1, ∞, that the complex dilatation μ_t of Φ_t is f-invariant, and that Φ_t is a C^1-function of t for fixed z. Then:

(1) Function $f_t = \Phi_t \circ f \circ \Phi_t^{-1}$ belongs to R_q for $t \in T$.

(2) There exists an arc $\alpha \subset T$ such that $f_t = f_s$ for $s, t \in \alpha$.

Proof

Since Φ_t fixes 0, 1, ∞, it follows that f_t is not defined at $\{0,\infty\}$, it does not take these values, and it maps C^* onto C^*. f is a local homeomorphism, so by *Lemma 4* f_t is locally conformal and has removable singularities besides 0 and ∞ which are essential singularities of f_t. Thus $f_t(z) = z^k \exp[\bar{F}(z) + \bar{G}(1/z)]$, where $\bar{F}$ and $\bar{G}$ are entire nonconstant and $k > 0$. If $a_1, \ldots, a_q$ are basic points of f, then

$$f\colon C^* - f^{-1}(\{a_1, \ldots, a_q\}) \to C^* - \{a_1, \ldots, a_q\}$$

is a covering map. Since f_t is topologically conjugated with f, then f_t has a finite set of basic points, i.e., $\Phi_t(a_1), \ldots, \Phi_t(a_q)$ and

$$f_t : C^* - \Phi_t \circ f^{-1}(\{a_1, \ldots, a_q\}) \to C^* - \{\Phi_t(a_1), \ldots, \Phi_t(a_q)\}$$

is also a covering map. We define a function k, $k(t) = [\Phi_t(a_1), \ldots, \Phi_t(a_q)]$ for $t \in T$; k is a C^1 map from $T \subset R^{2q+1}$ into $\mathbb{C}^q$. By a well-known topological theorem, there exists an arc $\alpha \subset T$ such that $\Phi_t(a_i) \equiv \bar{a}_i$, $i = 1, \ldots, 2q+1$, for $t \in \alpha$. Now we show that, if $r, s \in \alpha$, then $f_r = f_s$. Since

$$f_t : C^* - \Phi_t \circ f^{-1}(\{a_1, \ldots, a_q\}) \to C^* - \{\bar{a}_1, \ldots, \bar{a}_q\}$$

is a covering map, by the Covering Homotopy Theorem there exists a family of homeomorphisms

$$H_t\colon C^* - \Phi_r \circ f^{-1}(\{a_1, \ldots, a_q\}) \to C^* - \Phi_s \circ f^{-1}(\{a_1, \ldots, a_q\})$$

such that $\Phi_s \circ \Phi_t^{-1} \circ f_r = f_s \circ H_t$ for fixed $s, r \in \alpha$ and any $t \in \alpha$, and for $t = r$ we have $H_r = \Phi_s \circ \Phi_r^{-1}$. We may assume that 1 is a fixed point of f. It is a consequence of a theorem proved by Battacharyya (1969) that for any $n \in N$ there exist infinitely many periodic points of f with period n. Then $f_t(1) = 1$ for $t \in T$. This implies that $H_t(1) = 1$. If $t = s$, then $f_r = f_s \circ H_s$. Thus $\Phi_r \circ f \circ \Phi_r^{-1} = \Phi_s \circ f \circ \Phi_s^{-1} \circ H_s$ and $\Phi_s^{-1} \circ \Phi_r \circ f = f \circ \Phi_s^{-1} \circ H_s \circ \Phi_r$. On the other hand, we have $\Phi_s^{-1} \circ \Phi_r \circ f = f \circ \Phi_s^{-1} \circ \Phi_r$ since f_r and f_s are topologically conjugated. So $f \circ \Phi_s^{-1} \circ H_s \circ \Phi_r = f \circ \Phi_s \circ \Phi_r$ and $f \circ \Phi_s^{-1} \circ H_s = f \circ \Phi_s^{-1}$. Let $g = f \circ \Phi_s^{-1}$: then $g \circ H_s = g$, $g \circ H_s^2 = g$. We show that $H_s^2 =$ identity (id). This implies that $H_s =$ id or $H_s(z) = 1/z$. The last case is impossible in view of the continuity of H_s as a function of s. We note first that 0 is not a point of essential singularity of H_s, since in the neighborhood of this point H_s is not one-to-one. Suppose that 0 is a

removable singularity of H_s. Then there exists $\lim_{z \to 0} H(z) = a$. Since H_s is an open map of C^* onto C^*, an extension $\bar{H}_s$ of H_s given by $\bar{H}_s(0) = a$ is also an open map. Therefore for $a \in \mathbb{C}^*$ there exists a small enough neighborhood U of some point $b \in \mathbb{C}^*$ such that $f(U) \ni a$. But, if $\lim_{z \to 0} f(z) = \lim_{\omega = f(z)} \omega = a$, then the preimage of ω U belongs to some neighborhood of 0 and the function H_s is not one-to-one. Suppose that $\bar{H}_s(0) = 0$; then $\bar{H}_s \mid \mathbb{C}^* = H_s$. Then $\bar{H}_s$ is a holomorphic homeomorphism of $\mathbb{C}$ with two fixed points 0, 1. This implies that $H_s = \mathrm{id}$. If $\bar{H}_s(0) = \infty$, then $\bar{H}_s(\infty) = 0$. Otherwise we have only one possibility, i.e., $\bar{H}_s(\infty) = \infty$. Thus there exists a neighborhood U of 0 and neighborhood V of ∞ such that $\bar{H}_s(U)$ and $\bar{H}_s(V)$ are neighborhoods of ∞. This contradiction implies that, if $\bar{H}_s(0) = \infty$, then $\bar{H}_s(\infty) = 0$, so $\bar{H}_s^2(0) = 0$. It is clear that $\bar{H}_s^2$ is a holomorphic homeomorphism of $\mathbb{C}$ with two fixed points, so $\bar{H}_2 = \mathrm{id}$.

Lemma 7

Let D_0 be a component of $N(f)$ defined in *Proposition 3*, let t_0 be an end of the arc α defined in *Lemma 6*, and let $H_t = \Phi_{t_0}^{-1} \circ \Phi_t$ for $t \in \alpha$. Then $H_t = \mathrm{id}_{\mid J(f)}$ and $H_t(D_0) = D_0$.

Lemma 8

Let D_0, α, H_t be as in *Lemma 7*. If $g\colon B \to D_0$ is a conformal map from the unit disk B, then $h_t = g^{-1} \circ H_t \circ g$ extends to a homeomorphism $\bar{h}_t$ of $\mathrm{cl}\,B$ that satisfies $\bar{h}_{t \mid \partial B} = \mathrm{id}$.

Proof of Proposition 3

Suppose that D_0 is a wandering component of $f \in R_q$. Let g be a fixed conformal map from the unit disk B onto D_0, and let $\bar{a}, \bar{b}, \bar{c}$ be three distinct points of ∂B. Set

$$T = \{(t_1, \ldots, t_{2q+1}) : |t_i| < 1,\ t_i \in \mathbb{R},\ i = 1, \ldots, 2q+1\}$$

There exists a family φ_t, $t \in T$, of homeomorphism of $\mathrm{cl}B$ onto $\mathrm{cl}B$ such that $\varphi_{t \mid \partial B} \neq \varphi_{s \mid \partial B}$ for $t \neq s$, $\varphi_0 = \mathrm{id}$, any φ_t fixes $\bar{a}$, $\bar{b}$, $\bar{c}$, φ_t is the identity on the boundary arc $\overline{bc}$, φ_t is quasiconformal in B, the complex dilatation μ_t of φ_t is a continuous function of t at any $z \in B$, and $|\mu_t| < 1$. An example of such a family φ_t is described in detail in, e.g., Baker (1984) or Baker and Rippon (1984). Let $\Psi_t = g \circ \varphi_t \circ g^{-1}$. Then Ψ_t is a quasiconformal homeomorphism of D_0 onto itself and its complex dilatation is given by $\mu_{\Psi_t} = \mu_{\varphi_t}(g^{-1})\overline{(g^{-1})'}/(g^{-1})'$, which is continuous and thus measurable in z in D_0 and $|\mu_{\Psi_t}| < 1$. We extend μ_{Ψ_t} to an f-invariant complex dilatation in the plane (denoted by the same symbol) using the following construction. For $z \in \mathbb{C}^*$ such that there exist a z_1 in D_0 and positive integers n, m that

satisfy $f^n(z) = f^m(z_1)$, we set

$$\mu_{\Psi_t}(z) = \overline{(f^n)'(z)}\,(f^m)'(z_1)\mu_{\Psi_t}(z_1)/(f^n)'(z)\overline{(f^m)'(z_1)}$$

For other values of z we have $\mu_{\Psi_t} = 0$. We ought to check that μ_{Ψ_t} satisfies the differentiability conditions of *Lemma 5*. But the family φ_t described by Baker satisfies this condition. Then, by *Lemma 5*, there exists a family of quasiconformal homeomorphisms φ_t of $\bar{\mathbb{C}}$, $t \in T$, that satisfies the assumptions of *Lemma 6*, which implies that $f_t = \Phi_t \circ f \circ \Phi_t^{-1} \in R_q$ for $t \in T$ and $f_t = f_s$ for $t, s \in \alpha$, where α is some arc in T. Suppose that t_0 is an end of the arc α, $H_t = \Phi_{t_0}^{-1} \circ \Phi_t$, and $h_t = g^{-1} \circ H_t \circ g$ for $t \in \alpha$. *Lemma 3* (1) implies that $\mu_{h_t} = \mu_{H_t \cdot g}$. Since $\mu_{\Phi_t} = \mu_{\Psi_t}$ in B, $\Psi_t = g \circ \varphi_t \circ g^{-1}$, it follows that

$$\mu_{h_t} = \mu_{g \cdot \varphi_{t_0}^{-1} \cdot \varphi_t \cdot g^{-1} \cdot g} = \mu_{g \cdot \varphi_{t_0}^{-1} \cdot \varphi_t}$$

so by *Lemma 3* (1) it is the same as $\mu_{\varphi_{t_0}^{-1} \cdot \varphi_t}$. *Lemma 3* (3) applied to μ_{h_t} and $\mu_{\varphi_t^{-1} \cdot \varphi_t}$ implies that $h_t = L_t(\varphi_{t_0^{-1}} \circ \varphi_t)$, where L_t is a conformal Möbius transformation of B. By *Lemma 8*, h_t extends to a homeomorphism $\bar{h}_t$ of cl B that satisfies $\bar{h}_{t\,|\,\partial B} = \mathrm{id}$, so $L_t(\varphi_{t_0^{-1}} \circ \varphi_t) = L_s(\varphi_{t_0^{-1}} \circ \varphi_s)$ for $t, s \in \alpha$. But φ_t is an identify on the boundary arc $\overline{bc}$ for $t \in T$; thus $L_t = L_s$ for $t, s \in \alpha$. As a consequence, $\varphi_t = \varphi_s$ for $t, s \in \alpha \subset T$. This contradiction implies the thesis of *Proposition 3*.

2.4. Conclusion

Suppose that there exists a wandering component D of $N(f)$ for $f \in R_q$. Let $\{D_n\}_0^\infty$ denote an orbit of D_0. By *Proposition 1* there exists an $n_0 \in N$ such that one of the following cases is satisfied:

(1) D_n has finite topological type and $f\colon D_n \to D_{n+1}$ is a homeomorphism for $n > n_0$.

(2) D_n is an annulus for $n > n_0$ and there exists a subsequence $\{D_{n_k}\}_1^\infty$ such that the degree of $f^{i_k}\colon D_{n_k} \to D_{n_{k+1}}$ is greater than 1 for $n_k > n_0$.

(3) D_n has infinite topological type for $n > n_0$.

Case (1) was eliminated in *Proposition 3*, case (2) in *Proposition 2*. In the last case, i.e., if D_n has infinite topological type, the proof is analogous to Sullivan's (the dimension of the Teichmüler space of conformal structures for the direct limit D_∞ of $\{D_n\}_1^\infty$ is infinite). Thus we have proved the following theorem.

Theorem 1

For $f \in R_q$ every component of $N(f)$ is nonwandering.

3. Properties Dynamics for $f \in R_q$

3.1. Classification of components of $N(f)$

First we formulate one remark concerning dynamics in the simply connected components of a Radström function, i.e., $f \in R$.

Lemma 9

Let D be a simply connected periodic component of $N(f)$ for $f \in R$. Then D is one of the following components:

(1) D is an attracting (superattracting) domain.
(2) D is a parabolic domain.
(3) D is a Siegel disk.
(4) D is a trajectory $\{f^n(z)\}_1^\infty$ that escapes to ∞ for any $z \in D$.
(5) D is a trajectory $\{f^n(z)\}_1^\infty$ that escapes to 0 for any $z \in D$.

For rational maps any simply connected periodic domain is such as in (1), (2), (3); for entire transcendental maps there exists one more, i.e., (4). In this case there exists also a domain as described in (5). Of course, the proof of this lemma is similar to the one given for entire functions. It is enough to consider in that proof a limit function $f_0 \equiv 0$ instead of $f_0 \equiv \infty$.

Proposition 4

Let $f \in R_q$, let r_1, r_2 be radii such that an annulus $A = \{z \in \mathbb{C}^*: r_1 < |z| < r_2\}$ contains all basic points of f, and let H_i, $i = 1,2$, be components of $C^* - A$. Then $f: G_i = f^{-1}(H_i) \to H_i$ is a universal covering map and every component $V \in G_i$ is a simply connected domain whose boundary is a curve that tends in both directions either to 0 or to ∞.

We omit the proof of this proposition. Now we describe logarithmic change of coordinates in the neighborhood of 0 and ∞. Let H_i, G_i, $i = 1, 2$, be as in *Proposition 4*. Since $0 \notin G_i, H_i$, we may define $U_i = \ln G_i$, $P_1 = \ln H_1 = \{w: \operatorname{Re} w < \ln r_1\}$, and $P_2 = \{w: \operatorname{Re} w > \ln r_2\}$, and there exist maps $F_i: U_i \to P_i$ such that the following diagrams commutate:

$$\begin{array}{ccc} U_i & \xrightarrow{F_i} & P_i \\ {\scriptstyle \exp}\downarrow & & \downarrow{\scriptstyle \exp} \\ G_i & \xrightarrow[f]{} & H_i \end{array} \qquad (1)$$

Of course, the F_i are conformal univalent functions on every component of U_i, $i = 1, 2$. We may assume that $0 < r_1 < 1$ and $r_2 > 1$. Let W_i be a component of U_i, and $I_i : P_i \to W_i$ be a branch of the inverse function F_i^{-1}. For

$w \in W_1$, a ball at w with a radius $s = \ln r_1 - \mathrm{Re}F_1(w)$ is contained in the half-plane P_1. If $w \in W_2$, a ball at w with a radius $s = \mathrm{Re}F_2(w) - \ln r_2$ is contained in the half-plane P_2. By Koebe's theorem, each ball contains (respectively) a ball with a radius

$$(1/4) \mid I_1[F_1(w)] \mid [\ln r_1 - \mathrm{Re}F_1(w)]$$

or

$$(1/4) \mid I_2[F_2(w)] \mid [\mathrm{Re}F_2(w) - \ln r_2]$$

But the exponential function is one-to-one on W_i, so W_i does not contain vertical intervals with length greater than 2π. This implies that

$$|F'_1(w)| \geq (1/4\pi)[\ln r_1 - \mathrm{Re}F_1(w)] \qquad \text{for} \qquad w \in W_1 \tag{2}$$

$$|F'_2(w)| \geq (1/4\pi)[\mathrm{Re}F_2(w) - \ln r_1] \qquad \text{for} \qquad w \in W_2 \tag{3}$$

Proposition 5

For $f \in R_q$ there do not exist components of type (4) or (5) described in *Lemma 9*.

Proof

It is enough to prove that components of type (5) do not exist.

The proof in case (4) is the same as that for entire maps. Suppose that there exists a periodic component D such that $f^n(z) \to 0$ for all $z \in D$. We may assume that $f(D) = D$. Let $B(z,d)$ be a ball at z with radius d contained in D. Then the family $\{f^n \mid B\}_1^\infty$ is uniformly convergent, $f_0 \equiv 0$, and $B_n = f^n(B) \subset D$ for $n \in N$. If A is a component of $\ln B$, $A_n = F_1^n(A)$, then $\exp(A_n) = B_n$, $A_n \subset U_1$, and $\mathrm{Re}F_1(w)$ tends uniformly to $-\infty$ at A. Set $w \in A$, $W_n = F^n(w) \in A_n$, and let d_n be a radius of a maximal ball at w_n contained in A_n. By Koebe's theorem, $d_{n+1} \geq (1/4)d_n |F'_1(w_n)|$. Since $\mathrm{Re}F_1(w_n) \to -\infty$, by equation (2) $|F'(w_n)| \to \infty$. This implies that $d_n \to \infty$ and $A_n \subset U_1$ contains a vertical interval with length greater than 2π. This contradiction proves the thesis of the proposition.

As a consequence of Sullivan's theorem for $f \in R_q$ and *Proposition 5* we have a full classification of periodic components of $N(f)$.

Theorem 2

Let D be a periodic component of $N(f)$ for $f \in R_q$. Then D is either a Fatou

domain (i.e., attracting, superattracting, parabolic) or a rotation domain (i.e., Siegel disk, Herman ring).

Lemma 10

(1) If $\{D_i\}_{i=0}^{n-1}$ is a Fatou domain of $f \in R_q$, then f is not univalent in any D_i and $f(D_i)$ contains a basic point of f.
(2) If $\{D_i\}_{i=0}^{n-1}$ is a rotation domain of $f \in R_q$, then $\partial D \subset \mathrm{cl}[\bigcup_{n \in N} f^n(C)]$ where C denotes the set of basic points of f.

This result was obtained by Fatou (1919) for rational maps and may be extended without any change for entire or Radström functions.

Proposition 6

Let $f \in R_q$: then f has at most q Fatou domains and at most $2q$ Siegel disks.

One proof of this estimation for entire functions is based on the results from Nevanlinna's theory for these functions, e.g., Nevanlinna's second fundamental theorem. But these results are valid for Radström functions (see Battacharyya, 1969). Consequently, this proposition may be proved like Theorem 2 in the paper of Eremenko and Ljubič (1984a).

3.2. Estimation of the Lebesgue measure of $J(f)$

We give some estimation of the Lebesgue measure of the Julia set for $f \in R_q$ that satisfies some additional assumptions. We introduce the following notations. Let a be a basic point of f and a be an asymptotic value of f. Let $\{z \in C^* : |z - a| < \varepsilon\}$ be a ball which does not contain other basic points of f, and let W be a component of the set $\{z: |f(z) - a| < \varepsilon\}$ such that either $\mathrm{cl}W \cap \{0\} \neq \phi$ or $\mathrm{cl}W \cap \{\infty\} \neq \phi$. Of course, W is a simply connected domain. Let $\mathcal{W} = \ln W$ and $L = \bigcup_{k=-\infty}^{\infty}(L_0 + 2k\pi i)$, where L_0 is a component of L. Then L_0 is an unbounded strip which does not contain vertical intervals with length greater than 2π. By $O(x)$ we denote the total length of $L_0 \cap \{z: \mathrm{Re}\, z = x\}$. If $f \in R$ and $\ln\ln M(r,f) = O(\ln r)$ for $r \to 0$ and $r \to \infty$, where $M(r,f) = \max_{|z|=r} |f(z)|$, then there exist $M > 0$ and t_1, t_2 such that

$$\int_{t}^{t_1} [O(x)]^{-1}\,dx \leq -Mt \qquad \text{for} \qquad -\infty < t < t_1 < 0 \tag{4}$$

$$\int_{t_2}^{t} [O(x)]^{-1}\,dx \leq Mt \qquad \text{for} \qquad 0 < t_2 < t < \infty \tag{5}$$

Equation (4) is true if a is an asymptotic value along the path $\gamma \to 0$, while equation (5) is true if $\gamma \to \infty$. The estimation (5) has been proved by Ahlfors (see Nevanlinna, 1970) for entire functions, but it is not difficult to extend it for $f \in R$.

Proposition 7

Let $f \in R_q$, and $\ln \ln M(r,f) = O(\ln r)$ for $r \to 0$ and $r \to \infty$, and suppose that at least one basic point a of f is an asymptotic value. Then there exist M_1, M_2 that satisfy $M_1 \leq \lim_{n \to \infty} |f^n(z)| \leq M_2$.

To prove this proposition we need the following lemmas.

Lemma 11

Let $f \in R$ and $\ln \ln M(r,f) = O(\ln r)$ for $r \to 0$ and $r \to \infty$. There exist $k < 1$, $\alpha > 0$, $t_1 < 0$, $t_2 > 0$ such that, if $R = \operatorname{Re} z$, $R < t_1$, or $R > t_2$, then $\operatorname{mes}[B(z,kR) \cap L] / \operatorname{mes}[B(z,kR)] > \alpha$.

Proof

By equations (4) and (5) there exists an M, $t_1 < 0$, $t_2 > 0$, such that

$$\int_t^{t_1} [O(x)]^{-1} \, dx < -Mt \qquad \text{for} \qquad -\infty < t < t_1 \tag{6}$$

$$\int_{t_2}^{t} [O(x)]^{-1} \, dx < Mt \qquad \text{for} \qquad t_2 < t < \infty \tag{7}$$

Applying the Schwartz inequality to the inequalities (6) and (7), we have

$$\int_t^{t_1} [O(x)]^{-1} \, dx \int_{t_2}^{t_1} O(x) dx \geq (t_1 - t)^2 \qquad \text{for} \qquad t < t_1$$

and

$$\int_{t_2}^{t} [O(x)]^{-1} \, dx \int_{t_2}^{t_1} O(x) dx \geq (t_1 - t_2)^2 \qquad \text{for} \qquad t > t_2$$

This implies that $\int_t^{t_1} O(x) dx \geq -M^{-1}t$ for $t < t_1$ and $\int_{t_2}^{t} O(x) dx \geq M^{-1}t$ for $t > t_2$. Since $O(x) \leq 2\pi$, it follows that, for small enough k, $2k\pi < M^{-1}$ and

$$\int_{t}^{t_1} O(x)dx \geq -M^{-1}t + 2k\pi t \geq -\eta t \qquad \text{for} \qquad t < t_1$$

$$\int_{t_2}^{t} O(x)dx \geq M^{-1}t - 2k\pi t \geq \eta t \qquad \text{for} \qquad t > t_2$$

Thus $\operatorname{mes}[B(z,kR) \cap L]/\operatorname{mes}[B(z,kR)] \geq \alpha > 0$ for $\operatorname{Re} z < t_1$ and $\operatorname{Re} z > t_2$. Now we recall Koebe's theorem.

Lemma 12

Let g be a univalent holomorphic function defined in $B(z,R)$, $k < 1$:

(1) If $|w - z| = R$, then

$$|g'(z)|k/(1+k)^2 \leq |g(w) - g(z)| \leq |g'(z)|k/(1-k)^2$$

(2) If $|z| < kR$ and $|w| < kR$, then $|g'(z)/g'(w)|' < T(k)$.

Proof of Proposition 7

First we prove that there exists an M_1 such that $M_1 \leq \underline{\lim}\, |f^n(z)|$ for a.e. $z \in \mathbb{C}^*$. Suppose that a is an asymptotic value of f along the path $\gamma \to 0$, $a \neq 0, \infty$. We recall equation (1):

$$\begin{array}{ccccc} & & U_1 & \xrightarrow{F_1} & P_1 = \{w: \operatorname{Re} w < \ln r_1 < 0\} \\ & \exp & \downarrow & & \downarrow \exp \\ & & G_1 & \xrightarrow[f]{} & H_1 = \{z: |z| < r_1 < 1\} \end{array}$$

Let c be a constant that satisfies:

(1) $\exp c < |a| - \varepsilon < r_1$ for some $\varepsilon > 0$.
(2) $c < t_1$, where t_1 is defined in *Lemma 11*.
(3) If k is as in *Lemma 11*, then $-c > 2\pi(1+k)^2/(1-k)^2 = 2d$.
(4) Let $s > 1$: then $\operatorname{Re} F_1(w) < c$ if $|F'_1(w)| > 4s$.

Set $Y = \{w: \operatorname{Re} F_1^m(w) < c,\ n = 0,1,\ldots\}$. We shall prove that $\operatorname{mes} Y = 0$. Thus it is enough to prove that the density of the Lebesgue measure at any point $w \in Y$ is smaller than 1. Let $w \in Y$, $w_k = F_1^k w$, and $F_1^{-1}: P_1 \to U_1$ be a branch of the inverse function that satisfies $F^{-1}_{w_k = w_{k+1}}$. Since the image of U under F_1^{-1} does not contain vertical intervals with length greater than 2π, by *Lemma 12*,

$$F_1^{-1}\,[B(w_n,kR_n)] \subset B(w_{n-1},d) \tag{9}$$

where $R_n = \operatorname{Re} w_n$ and $d = \pi(1+k)^2/(1-k)^2$. Let F_1^{-1} be a branch of the inverse function that satisfies $F^{-1}w_l = w_{l-1}$, $l \in [1,n-1]$. As a consequence of the conditions imposed on c, the function F_1^{-1} is defined in $B(w_l,2d)$ and $|(F_1^{-1})'(w)| \le 1/(4s)$. Hence by *Lemma 12*(2) we have

$$F_1^{-1}\,[B(w_l,d)] \subset B(w_{l-1},s^{-1}d) \tag{10}$$

The inequalities (9) and (10) imply

$$F^{-n}[B(w_n,kR_n)] \subset B(w,s^{-n+1}d) \tag{11}$$

Let $B_n(w) = F^{-n}B(w_n,kR_n)$. By *Lemma 12*(2) we have

$$B(w,tr_n) \subset B_n(w) \subset B(w,r_n) \tag{12}$$

where t does not depend on n. Here $B(w,r_n)$ denotes the smallest ball around $B_n(w)$. By the inclusion (11),

$$r_n \le s^{-n+1}d \to 0 \qquad n \to \infty \tag{13}$$

Lemma 11 implies that $\operatorname{mes}[B(w,kR_n) \cap L]/\operatorname{mes}[B(w,kR_n)] \ge \alpha$. Applying *Lemma 12*(2) we have

$$\operatorname{mes}[B_n(w) \cap F^{-n}L]/\operatorname{mes}[B_n(w) \ge [T(k)]^{-2}\alpha \tag{14}$$

If $z \in F^{-n}L$, then $|\exp[F^{n+1}(z) - \alpha]| < \varepsilon$; but, if $z \in Y$, then $|\exp[F^{n+1}(z)]| < \exp c < |\alpha| - \varepsilon$. Hence $Y \cap f^{-n}L = \phi$ and equation (14) implies that

$$\operatorname{mes}[B_n(w) \cap L]/\operatorname{mes}[B_n(w)] \le 1 - [T(k)]^{-2}\alpha$$

This means that the density of the Lebesgue measure of Y at W is smaller than 1.

If α is an asymptotic value of f along the path $\gamma \to 0$, then we consider function $\exp(-w)$ in diagram (8) instead of $\exp(w)$ and $P_2 = \{w\colon \operatorname{Re} w > \max(-\ln r_1, \ln r_2)\}$. As a consequence, it is necessary to change conditions (1)–(4) imposed on c. It remains to prove that there exists a constant M_2 such that $\underline{\lim}\,|f^n(z)| \le M_2$ for a.e. $z \in \mathbb{C}^*$. If α is an asymptotic value along the path $\gamma \to \infty$, then we repeat the proof for entire

functions. For an asymptotic value along $\gamma \to 0$, we consider $\ln z^{-1}$ instead of $\ln z$.

Theorem 3

Let f be as in *Proposition 7*. If $J(f) \neq C^*$ and for any basic point a_k of f which belongs to $J(f)$ there exists an $n_k \in N$ and a repelling periodic point b_k such that $b_k = f^{n_k}(a_k)$, then $\mathrm{mes}[J(f)] = 0$.

Proof

Let $\{b_n\}$, $k = 1, \ldots, l$, denote all repelling periodic points such that $b_k = f^{n_k}(a_k)$ for some $n \in N$ and basic point a_k, and let $\{c_i\}, i = 1, \ldots, n$, denote all rational elliptic points of f. By *Proposition 7* there exist M_1, M_2 such that

$$2 \max_{k,i} |b_k|, |c_i| < M_2 \text{ and } \underline{\lim} |f^n(z)| < M_2 \quad \text{for a.e.} \quad z \in \mathbb{C}^*$$

$$M_1 < \min_{k,i} |b_k|, |c_i| \text{ and } M_1 < \underline{\lim} |f^n(z)| \quad \text{for a.e.} \quad z \in \mathbb{C}^*$$

We choose a point $z \in J(f)$ such that $M_1 < \underline{\lim} |f^n(z)| < M_2$ and there exists neither a repelling periodic point a nor an elliptic rational point a satisfying $f^n(z) = a$ for some $n \in N$. Thus there exist a $\delta > 0$ and a sequence $\{n_s\}_{s=1}^{\infty}$ such that

$$M_1 < |f^{n_s}(z)| < M_2 \quad \inf_{k,i}(f^{n_s} - b_k, |f^{n_s} - c_i|) > \delta$$

Hence we may define f^{-n_s} on $B(f^{n_s}z, \delta)$. Since $J(f) \neq \mathbb{C}^*$, it follows that

$$\inf_{M_1 < |w| < M_2} \{\mathrm{mes}[B(w, \delta/2) \cap N(f)] / \mathrm{mes}[B(w, \delta/2)]\} \geq \alpha > 0$$

By *Lemma 12*(2), $\mathrm{mes}[B_{n_s}(w) \cap N(f)] / \mathrm{mes}[B_{n_s}(w)] \geq [T(1/2)]^{-1} \alpha$, where $B_{n_s}(w) = f^{-n_s} B[f^{n_s}(w), \delta/2]$. This implies that the density of the Lebesgue measure of $J(f)$ at w is smaller than 1.

3.3. J-Stability for $f \in R_q$

The functions f and g are conjugated if there exist homeomorphisms φ and ψ of $\mathbb{C}^*$ such that $f \circ \varphi = \psi \circ g$. Let M_f denote the space of functions $f \in R$ conjugated with g. If $M = \{g \in M_f : f \in R_q\}$, then M is a manifold with a topology which is locally equivalent to the uniform topology on compact

subsets of $\mathbb{C}^*$. Let $\alpha_1(g), \ldots, \alpha_q(g)$ denote the basic points of $g \in R_q$. We choose $\alpha, \beta \in \mathbb{C}^*$, and by $M_f^{\alpha,\beta}$ we denote the space of functions $g \in M_f$ for which the conjugacy homeomorphisms φ and ψ fix α, β. Let $\alpha_{q+1}(g) = g(\alpha)$ and $\alpha_{q+2}(g) = g(\beta)$ if $\alpha, \beta \neq 0$; for $\alpha = 0$ or $\beta = 0$ we set $\alpha_{q+1}(g) = 0$ or $\alpha_{q+2}(g) = 0$. Then $M_f^{\alpha,\beta}$ is an analytic manifold with local coordinates $\{\alpha_i(g)\}_{i=1}^{q+2}$. If $M_f = \bigcup_{\alpha,\beta} = M_f^{\alpha,\beta}$, then M_f is an analytic manifold with topology locally equivalent to the uniform topology on a compact subset of $\mathbb{C}^*$. But it is more convenient to work with the factor space $\bar{M}$ given by the relation of conjugacy. Then $\bar{M}$ is a q-dimensional manifold with coordinates $[\alpha_1(f), \ldots, \alpha_q(f)]$. A function $f \in M$ is J-stable if for every $g \in M$ close enough to f there exists a homeomorphism $h: J(f) \to J(g)$ such that $f \circ h = h \circ g$.

The following lemma and theorem are analogous to those of Eremenko and Ljubič (1984b), but it is necessary to change their proof by considering asymptotic paths that tend to zero.

Lemma 13

A periodic point is a holomorphic map of $f \in M$ in the local coordinates on M that admits only algebraic singularities. An eigenvalue of the periodic point is a nonconstant holomorphic function of $f \in M$.

Theorem 4

J-stability is a generic property in M.

Conjecture

Structural stability is a generic property in M.

Acknowledgments

The author would like to thank R. Kopiecki and A. Zdunik of the Institute of Mathematics, Warsaw University, for informative conversations; she would also especially like to thank F. Przytycki of the Institute of Mathematics of the Polish Academy of Sciences with whom she worked during his stay at that institute in 1984–1985.

References

Ahlfors, L. and Bers, L. (1960), Riemann's mapping theorem for variable metrics, *Ann. of Math.*, **72**, 385–404.

Baker, I. N. (1968), Repulsive fixed points of entire functions, *Math. Z.*, **104**, 252–256.

Baker, I. N. (1976), An entire function which has wandering domains, *J. Austral. Math. Soc., Ser. A.*, **22**, 173–176.

Baker, I. N. (1984), Wandering domains in the iteration of entire functions, *Proc. London. Math. Soc.*, **49**, 563–576.

Baker, I. N. and Rippon, P. J. (1984), Iteration of exponential functions, *Ann. Acad. Sci. Fenn. Ser. A I Math.*, **9**, 49–77.

Battacharyya, P. (1969), *Iteration of Analytic Functions*, PhD thesis (University of London).

Eremenko, A. E. and Ljubič, M. Yu. (1984a), *Iterates of Entire Functions*, Preprint 6-84 (Physico-Technical Institute of Low Temperatures, Ukrainian SSR Academy of Sciences, Kharkov, USSR). (In Russian.)

Eremenko, A. E. and Ljubič, M. Yu. (1984b), *Structural Stability in Some Families of Entire Functions*, Preprint 29-84 (Physico-Technical Institute of Low Temperatures, Ukrainian SSR Academy of Sciences, Kharkov, USSR). (In Russian.)

Fatou, P. (1919), Sur les equations fonctionnelles, *Bull. Soc. Math. France*, **47**, 161–271.

Fatou, P. (1920a), Sur les equations fonctionnelles, *Bull. Soc. Math. France*, **48**, 33–94.

Fatou, P. (1920b), Sur les equations fonctionnelles, *Bull. Soc. Math. France*, **48**, 208–314.

Fatou, P. (1926), Sur l'itération des fonctions transcendantes entières, *Acta Math.*, **47**, 337–370.

Goldberg, L. and Keen, L. (to appear), A finiteness theorem for a dynamical class of entire functions.

Gol'dberg, A. A. and Ostrovskij, I. V. (1970), *Raspredelenie značenij meromorfnych funkcjij*, (Izdat. Nauka, Moscow, USSR). (In Russian.)

Julia, G. (1918), Memoire sur l'itération des fonctions rationnelles, *J. Math. Pures Appl.*, **8**, 47–245.

Lehto, O. and Virtanen, K. I. (1965), *Quasiconforme Abbildungen* (Springer, Berlin).

Mane, R., Sad, P., and Sullivan, D. (1983), On the dynamics of rational functions, *Ann. Sci. Ecole Norm. Sup.*, **16**, 193–217.

Nevanlinna, R. (1936), *Eindeutige analytische Funktionen* (Springer, Berlin).

Nevanlinna, R. (1970), *Analytic Functions* (Springer, Berlin).

Radström, H. (1953), On the itération of analytic functions, *Math. Scand.*, **1** 85–92.

Sullivan, D. (1982a), Itération des fonctions analytiques complex, *C.R. Acad. Sci. Paris. Ser. I Math.*, **294**(9), 301–303.

Sullivan, D. (1982b), *Quasiconformal Homeomorphisms and Dynamics I: Solution of the Fatou–Julia Problem of Wandering Domains*, Preprint M/82/59 (Institute des Hautes Etudes Scientifiques, Paris).

Sullivan, D. (1983), *Quasiconformal Homeomorphisms and Dynamics III: Topological Conjugacy Classes of Analytic Endomorphisms*, Preprint M/83/1 (Institute des Hautes Etudes Scientifiques, Paris).

II. VIABILITY THEORY AND MULTIVALUED DYNAMICS

A Viability Approach to Ljapunov's Second Method

J.-P. Aubin and H. Frankowska
VER Mathématiques de la Decision, Université de Paris IX — Dauphne, Place du Maréchal de Lattre de Tassigny, Paris 16, France

When f is a continuous single-valued map from an open subset Ω of $\mathbf{R}^n$ to $\mathbf{R}^n$ and V is a differentiable function defined on Ω, the Ljapunov method derives from estimates of the form

$$\forall x \in \Omega \qquad \langle V'(x), f(x) \rangle \leq \psi[V(x)] \tag{1}$$

information on the behavior of a solution $x(\cdot)$ to the differential equation $x' = f(x)$, $x(0) = x_0$, given by inequalities of the form

$$V[x(t)] \leq w(t) \tag{2}$$

where w is a solution to the differential equation

$$w'(t) = \psi[w(t)] \qquad w(0) = V(x_0) \tag{3}$$

(see for instance Yoshizawa, 1966).

We shall extend this result when we replace the differential equation by a differential inclusion, when we require viability conditions and when we assume that V is only continuous (because "interesting" examples of functions V are derived from nondifferentiable norms, for instance). We look for solutions $x(\cdot)$ to, for almost all $t \in [0,T]$,

$$x'(t) \in F[x(t)] \qquad x(0) = x_0 \qquad \text{given in } K \tag{4}$$

satisfying

$$\forall t \in [0, T] \qquad x(t) \text{ belongs to a closed subset } K \text{ (viability)} \tag{5a}$$

$$\forall t \in [0, T] \qquad V[x(t)] \leq w(t) \tag{5b}$$

where $w(t)$ is a solution to differential equation (3). For this purpose, we choose from the concepts of tangent cones to subsets and generalized directional derivatives of a function the *contingent cone* $T_K(x)$ to K at x, defined by

$$T_K(x) := \left\{ v \in \mathbf{R}^n \mid \liminf_{h \to 0+} \frac{d(x + hv, K)}{h} = 0 \right\} \tag{6}$$

and introduced by Bouligand (1932) (see also Aubin and Cellina, 1984, Section 4.2, pp. 176–177), and the *hypo-contingent derivative* $D\_V(x)$ of V at x, defined by

$$\forall v \in \mathbf{R}^n \qquad D - V(x)(v) := \limsup_{\substack{h \to 0+ \\ v' \to v}} \frac{V(x + hv') - V(x)}{h} \tag{7}$$

(Aubin and Cellina, 1984, Section 6.1, p. 287).

We shall prove the following.

Theorem 1

Let V be a nonnegative continuous function defined on a neighborhood of the closed subset K and ψ be a nonpositive continuous function from $\mathbf{R}_+$ to $\mathbf{R}$ satisfying $\psi(0) = 0$. Let $x_0 \in K$ be given.

(1) We assume that

$$F \text{ is upper semicontinuous with nonempty compact convex values} \tag{8}$$

If we replace estimate (1) by

$$\forall x \in K, \qquad \exists v \in F(x) \cap T_K(x) \text{ such that } D\_V(x)(v) \leq \psi[V(x)] \tag{9}$$

there exist $T > 0$ and solutions $w(\cdot)$, $x(\cdot)$ to the problem (3), (4), and (5).

(2) We assume that

$$F \text{ is continuous with nonempty compact values} \tag{10}$$

If we posit the stronger estimate

$$\forall x \in K, \qquad F(x) \subset T_K(x) \tag{11a}$$

and

$$\sup_{v \in F(x)} D\_V(x)(v) \leq \psi[V(x)] \tag{11b}$$

there exist $T > 0$ and solutions $w(\cdot)$, $x(\cdot)$ to the problem (3), (4), and (5).

(3) We assume that

$$\begin{aligned} &F \text{ is Lipschitz on a neighborhood of } K \text{ and has nonempty compact} \\ &\textit{values} \text{ and} \\ &\psi \text{ is Lipschitz on a neighborhood of } [0, w_0] \end{aligned} \tag{12}$$

Then estimate (11) implies the existence of $T > 0$ such that any solution $[w(\cdot), x(\cdot)]$ to (3) and (4) satisfies property (5).

Remark

If we assume furthermore that F is bounded, we can take $T = +\infty$ in the above theorem. This implies that $w(t)$ converges to some w_* when $t \to \infty$, where $w_* \in [0, V(x_0)]$ is a solution to the equation $\psi(w_*) = 0$. If $\psi(w) < 0$ for all $w > 0$, we then deduce that

$$\lim_{t \to \infty} V[x(t)] = 0 \tag{13}$$

Proof of Theorem 1

We set

$$G(x,w) := F(x) \times \psi(w) \subset \mathbf{R}^n \times \mathbf{R} \tag{14}$$

We introduce the viability domain

$$K := \{(x, w) \in K \times [0,w_0] \mid V(x) \leq w\} \tag{15}$$

which is a closed subset of $\mathbf{R}^n \times \mathbf{R} \times \mathbf{R}$ [where $w_0 > V(x_0)$]. We observe that if

$$v \in t_K(x) \qquad \text{satisfies} \qquad D\_V(x)(v) \leq \psi[V(x)] \tag{16}$$

then

$$[v, \psi(w)] \quad \text{belongs to} \quad T_K(x, w) \tag{17}$$

Indeed, since v belongs to $T_K(x)$, there exist sequences of elements $h_n > 0$ and v_n converging to zero and v such that

$$\forall n, \ x + h_n v_n \in K$$

By the very definition of $D\_V(x)(v)$, there exists a sequence of elements $a_n \in \mathbb{R}$ converging to $D\_V(x)(v)$, such that, for all $n \geq 0$,

$$V(x + h_n v_n) \leq V(x) + h_n a_n$$

If $\psi[V(x)] = w$, we take

$$b_n := a_n + \psi[V(x)] - D\_V(x)(w)$$

if $D\_V(x)(w) > -\infty$, and

$$b_n = \psi[V(x)]$$

if $D\_V(x)(w) = -\infty$. If $\psi[V(x)] < w$, we take $b_n := \psi(w)$ and we deduce that

$$V(x + h_n v_n) \leq w + h_n \psi(w)$$

for large enough n because V is continuous. In summary, b_n converges to w and satisfies

$$\forall n, \ x + h_n v_n \in K \quad \text{and} \quad V(x + h_n v_n) \leq w + h_n b_n \tag{18}$$

This shows that $(x + h_n v_n, w + h_n b_n)$ belongs to K and thus that $[v, \psi(w)]$ belongs to $T_K(x, w)$.

We consider now trajectories $x(\cdot)$, $w(\cdot)$ of the differential inclusion

$$[x'(t), w'(t)] \in G[x(t), w(t)] \tag{19a}$$

$$[x(0), w(0)] = [x_0, V(x_0)] \tag{19b}$$

which are viable in the sense that

$$\forall t \in [0, T] \qquad [x(t), w(t)] \in K \tag{20}$$

We then observe that $x(\cdot)$ is a solution to (4), that w is a solution to (3) and that (20) implies properties (5).

If F satisfies assumptions (8) and (9), then G is also upper semicontinuous with compact convex values and $G(x, w) \cap T_K(x, w) \neq \phi$. Hence Haddad's viability theorem (Haddad, 1981; Aubin and Cellina, 1984, p. 180, Theorem 4.2.1) implies the existence of a solution to (19) and (20) on some interval.

If F satisfies assumptions (10) and (11), then G is continuous with compact values and $G(x, w) \subset T_K(x, w)$. Hence the viability theorem of Aubin and Clarke (1977) (see also Aubin and Cellina, 1984, p. 198, Theorem 4.6.1) implies the existence of a solution to (19) and (20) on some interval.

If F satisfies assumptions (11) and (12), then G is Lipschitz with compact values on a neighborhood of K and $G(x, w) \subset T_K(x, w)$. Hence the invariance theorem of Clarke (1975) (see also Aubin and Cellina, 184, p. 202, Theorem 4.6.20) shows that any solution of (19) satisfies (20).

Remark

We can solve in the same way the case when we consider

$$\begin{array}{l} p \text{ nonnegative continuous functions } V_j \text{ defined} \\ \text{on a neighborhood of K} \end{array} \tag{21a}$$

$$\begin{array}{l} p \text{ nonpositive continuous functions } \psi_j \text{ from} \\ \mathbf{R}_+ \text{ to } \mathbf{R}_+ \textit{satisfying } \psi_j(0) = 0 \end{array} \tag{21b}$$

and when we replace condition (15) by

$$\forall t \in [0, T] \qquad x(t) \in K \tag{22a}$$

$$\forall t \in [0, T] \qquad \forall j = 1, \ldots, p \qquad V_j[x(t)] \leq w_j(t) \tag{22b}$$

where $w_j(\cdot)$ is some solution to the differential equation

$$w_j'(t) = \psi_j[w_j(t)] \qquad w_j(0) = V_j(x_0) \tag{23}$$

We have to replace the Ljapunov estimates (9) by

$$\begin{array}{l} \forall x \in K, \qquad \exists v \in F(x) \cap T_K(x) \qquad \text{such that} \\ \forall j = 1, \ldots, p, \qquad D\_V_j(x)(v) \leq \psi_j[V_j(x)] \end{array} \tag{24}$$

and estimates (11) by

$$\forall x \in K, \qquad F(x) \subset T_K(x) \qquad \text{and} \qquad \forall j = 1, \ldots, p$$
$$\sup_{v \in F(x)} D\_V_j(x)(v) \leq \psi_j[V_j(x)] \tag{25}$$

Therefore the asymptotic properties of solutions to the differential inclusions (4) are concealed in the function ψ_0 defined by

$$\psi_0(w) := \sup_{V(x)=w} \inf_{u \in F(x) \cap T_K(x)} D\_V(x)(u) \tag{26}$$

for set-valued maps F satisfying (8) and the function ψ_1 defined by

$$\psi_1(w) := \sup_{V(x)=w} \sup_{u \in F(x)} D\_V(x)(u) \tag{27}$$

for set-valued maps F satisfying (10). Hence any continuous function ψ larger than ψ_0 (or ψ_1) will provide solutions $w(\cdot)$ to (3) estimating the value $V[x(t)]$ on some trajectory of the differential inclusion (4).

For instance, we obtain the following consequence on asymptotic stability.

Corollary

Let V be a nonnegative continuous function defined on a neighborhood of K and let x_0 be given. Let F satisfy assumption (8). We assume further that $\lambda_0 \in \mathbf{R}$ achieves the finite maximum in

$$\rho_0 := \sup_{\lambda \in \mathbf{R}} \inf_{w \geq 0} [\lambda w - \psi_0(w)] \tag{28}$$

If $\rho_0 > 0$ and

$$V(x_0) \leq \frac{\rho_0}{\lambda_0} \{1 - \exp(-\lambda_0 T)\}$$

there exists a solution $x(\cdot)$ to the differential inclusion (4) satisfying

$$\forall t \in [0,T], \qquad V[x(t)] \leq \begin{cases} \dfrac{\rho_0}{\lambda_0} \{1 - \exp[\lambda_0(t-T)]\} & \text{if} \quad \lambda_0 \neq 0 \\ -\rho_0(t-T) \end{cases} \tag{29}$$

If $\rho_0 \leq 0$ and $\lambda_0 < 0$, then there exists a solution $x(\cdot)$ to the differential inclusion (4) satisfying

$$\forall t \geq 0, \qquad V[x(t)] \leq \frac{1}{\lambda_0}[\rho_0 - c_0 \exp(\lambda_0 t)] \tag{30}$$

where $c_0 = \rho_0 - \lambda_0 V(x_0)$.

Proof

We take $\psi(w) := \lambda_0 w - \rho_0$.

Remark

Theorem 1 implies directly the asymptotic properties on U-monotone maps as they appear in Corollaries 6.5.1 and 6.5.2 of Aubin and Cellina (1984, pp. 320–332).

Let $U : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}_+ \cup \{\infty\}$ be a nonnegative function satisfying

$$U(y, y) = 0 \qquad \text{for all} \qquad y \in K \tag{31}$$

which plays the role of a semidistance (without having to obey the triangle inequality).

We assume that, for all $x \in K$, $x \to U(x, y)$ is locally Lipschitz around K and we set

$$U'(x, y)(v) := D_-[x \to U(x, y)](x)(v) \tag{32}$$

Let Φ be a continuous map from $\mathbf{R}_+$ to $\mathbf{R}_+$ such that $\Phi(0) = 0$. We say that F is U-monotone (with respect to Φ) if

$$\begin{gathered} \forall x, \qquad y \in K, \qquad \forall u \in F(x), \qquad \forall v \in F(y), \\ U'(x, y)(v - u) + \Phi[U(x, y] \leq 0 \end{gathered} \tag{33}$$

Let us assume that $c \in K$ is an equilibrium of F [a solution to $0 \in F(c)$] and that $-F$ is U-monotone with respect to Φ. Then we observe that by taking $V(x) := U(x, c)$ we have

$$\psi_0(w) \leq \psi_1(w) \leq -\Phi(w) \tag{34}$$

Let $w(\cdot)$ be a solution to the differential equation

$$w'(t) + \varphi[w(t)] = 0 \qquad w(0) = U(x_0, c) \tag{35}$$

If F satisfies either (8) or (10), there exists a solution to the differential inclusion (4) satisfying

$$U[x(t), c] \leq w(t) \qquad \text{for all} \qquad t \in [0, T] \tag{36}$$

References

Aubin, J.P. and Cellina, A. (1984), *Differential Inclusions* (Springer, Berlin).

Aubin, J.P. and Clarke, F.H. (1977), Monotone invariant solutions to differential inclusions, *J. Lond. Math. Soc.*, **16**, 357–366.

Bouligand, G. (1932), *Introduction à la géométrie Infinitésimale Directe* (Gauther-Villars, Paris).

Clarke, F.H. (1975), Generalized gradients and applications, *Trans. Am. Math. Soc.*, **205**, 247–262.

Haddad, G. (1981), Monotone trajectories of differential inclusions and functional differential inclusions with memory, *Isr. J. Math.*, **39**, 83–100.

Yoshizawa, T. (1966), *Stability Theory by Liapunov's Second Method* (Mathematical Society of Japan, Tokyo).

Repellers for Generalized Semidynamical Systems

V. Hutson[1] and J.S. Pym[2]
[1]*Department of Applied Mathematics, The University, Sheffield, UK*
[2]*Department of Pure Mathematics, The University, Sheffield, UK*

1. Introduction

The motivation for the analysis to be presented comes from a problem that is one of the most fundamental in biology: to find the conditions under which a system of interacting species, genes, etc., are assured of "long-term survival". What exactly is meant by this term is itself not clear and has been discussed (see Hofbauer, 1985). We shall not repeat the arguments given there, but shall accept that for dynamical systems on $\mathbf{R}_n^+$, for example, generated by ordinary differential equations, a criterion that is biologically reasonable is that there should be a compact absorbing set M in the interior $\overset{0}{\mathbf{R}}{}_n^+$ of $\mathbf{R}_n^+$ for all semiorbits with initial values in $\overset{0}{\mathbf{R}}{}_n^+$, called *permanent coexistence* (or permanence). There is now a considerable amount known concerning mathematical techniques for tackling this question for systems governed by ordinary differential equations, difference equations, partial differential equations, and modeling spatial diffusion of species (see, e.g., Amann and Hofbauer, 1984; Hutson and Moran, 1982, 1985; Butler *et al.*, 1985).

Whilst much remains to be done in this area, there is another fundamental issue affecting the problem of long-term survival; namely, that with the present state of empirical knowledge, the basic model for a given biological system cannot realistically be regarded as known. Indeed, it may be a difficult matter to decide what type of equation provides a reasonable approximation. Even if in ecology, for example, the assumption is made that the governing equations are ordinary differential equations, say

$$\dot{x}_i = x_i f_i(x) \qquad (i = 1, \ldots, n) \tag{1}$$

it is clear that for only three species it would be a major experimental project to find a reasonable approximation to the f_i. It is therefore important to enquire whether, if the f_i are known only very roughly, anything can be said concerning the question of permanent coexistence, at least.

For ordinary differential equations a suitable setting for tackling this question is the theory of differential inclusions (Aubin and Cellina, 1984),

$$\dot{x}_i \in x_i F_i(x) \tag{2}$$

where the F_i are set-valued functions and the solutions are absolutely continuous functions satisfying system (2). These relations could be interpreted as differential inequalities into which the uncertainty concerning the measurements are subsumed, or they could be regarded as modeling unpredictable external factors, such as climate. Rather than treating system (2) directly, the analysis is carried out in terms of the theory of generalized semidynamical systems (GSDS), where the phase map π is set-valued. Such a system bears a relation to differential inclusions analogous to that which the standard dynamical system bears to differential equations. There are two reasons for using a GSDS: first, it is technically convenient; second, the added generality of the theory allows the possibility of treating more general models for biological interactions. For example, it is likely that similar results may be obtained for the semigroup $\mathbf{Z}^+$ as well as $\mathbf{R}^+$, which will allow "difference inclusions" (an analogous generalization of difference equations, see Laricheva, 1984) to be treated. We finally return to system (2) and show how the theory may be applied to a specific set of differential inclusions.

Suppose, then, that $\pi(x,t)$ represents the set of all possible values of solutions of system (2) through x at time t.

Definition 1

The system (2) will be said to be *permanently coexistent* if and only if there is a compact set $M \subset \overset{0}{\mathbf{R}}{}^+_n$ with the property that, given any $x \in \overset{0}{\mathbf{R}}{}^+_n$, there is a t_x such that $\pi(x,t) \subset M$ for $t \geq t_x$.

One can picture π as being represented by a "funnel". Because of the form of system (2), the boundary $\partial \mathbf{R}^+_n$ of $\mathbf{R}^+_n$ is an invariant set, so from the point of view of dynamical systems the aim is to find a practical method for discovering when an invariant set repels orbits in the strong sense of the definition.

2. Generalized Semidynamical Systems

These are also known as set-valued and multivalued dynamical systems, and have been studied by Barbashin (1948), Roxin (1965), Szegö and Treccani (1969), Kloeden (1979), and Dochev (1979), among others. They are a natural generalization of the usual semidynamical system, but various assumptions concerning continuity and backward extendability lead to a profusion of definitions. It turns out that in the present context continuity is unnecessarily restrictive and upper semicontinuity is enough.

Definition 2

Let X,Y be metric spaces and let $A(Y)$ denote the set of nonempty subsets of

Y. Then $F: X \to A(Y)$ is *upper semicontinuous* (USC) if and only if given any $x_0 \in X$ and any open $U \supset F(x_0)$, there is a neighborhood V of x_0 such that $F(x) \subset U$ for $x \in V$ (Aubin and Cellina, 1984).

With (X,d) a metric space, consider the map $\pi: X \times \mathbf{R}^+ \to A(X)$, and for $X_0 \subset X$, $I \subset \mathbf{R}^+$ put

$$\pi(X_0,I) = \bigcup_{x \in X_0} \bigcup_{t \in I} \pi(x,t)$$

It is convenient sometimes to write $\pi(x,t) = xt$ as usual. We shall assume that the GSDS $(X,\pi,\mathbf{R}^+)$ satisfies the following:

(1) $\pi(x,0) = x$.
(2) $\pi[\pi(x,t_1),t_2] = \pi(x,t_1 + t_2)$ $(t_1,t_2 \in \mathbf{R}^+)$.
(3) π is USC and compact valued.

Definition 3

$\gamma^+(x) = \{y: y \in \pi(x,t)$ for some $t \geq 0\}$ is the *semiorbit through* x, $\gamma^+(X_0)$ being defined by taking unions. X_0 is *forward invariant* if and only if $\gamma^+(X_0) \subset X_0$. The Ω-limit set $\Omega(X_0)$ of X_0 is the set

$$\Omega(X_0) = \bigcap_{t \geq 0} \mathrm{cl}\,[\bigcup_{x \in X_0} \gamma^+(xt)]$$

Definition 4

The set $M \subset X$ is said to be *absorbing* for X_0 if and only if given any $x \in X_0$, there exists a $t_x < \infty$ such that $xt \subset M$ for $t \geq t_x$. That is the "sections" xt of every semiorbit starting in X_0 are eventually contained in M.

Finally, to simplify the notation, for $P: X \to \mathbf{R}$ and $X_0 \subset X$ we write

$$\inf P(X_0) = \inf_{x \in X_0} P(x)$$

For the distance $d(S,M)$ between two sets we use the definition

$$d(S,M) = \inf_{x \in S} d(x,M)$$

3. Average Ljapunov Functions

In the biological context, because of the finiteness of the world we may assume that X is a compact neighborhood of the origin in $\mathbf{R}^+_n$. Then

$S = X \cap \partial \mathbb{R}^+_n$ is a compact, forward-invariant set. The conditions that will be given ensure that S repels orbits in a suitable sense, and are a weakening of the requirements imposed on the standard Ljapunov function (for repellers rather than attractors). The following discussion is intended to clarify the form of these conditions.

Temporarily, let π be an ordinary (not generalized) dynamical system. Let (the Ljapunov function) $P: X \to \mathbb{R}^+$ be continuous with $P^{-1}(0) = S$, and assume that on a neighborhood of S

$$P(xt)/P(x) > 1 \qquad (t > 0, x \in X \setminus S) \tag{3}$$

Then there is a neighborhood U of S with $X \setminus U$ absorbing for $X \setminus S$. The difficulty with this well-known result is familiar: it is usually hard to find a suitable P. For differential equations an ingenious result, which sometimes obviates this problem, was introduced by Schuster *et al.* (1979) and Hofbauer (1981). In a dynamical systems context (Hutson, 1984), it is enough to assume instead of system (3) that

$$\sup_{t>0} \liminf_{\substack{y \to x \\ y \in X \setminus S}} P(yt)/P(y) > 1 \qquad (x \in S) \tag{4}$$

This weakens the requirements in two ways. First, the condition need only hold on S. Second, it need only hold for *some* t, so that the value of P may first decrease along orbits, so long as at some stage it increases. This is quite a significant advance for, as we shall see later, it means roughly that the inequality need be checked only on $\Omega(S)$, which is often a much easier task; this is plausible as we can take a large t when the orbit is near $\Omega(S)$. P might be called an average Ljapunov function. From another point of view, it is at least not unreasonable that condition (4) should ensure that orbits are pushed away from S. The real point is that they may not return and come "too close" to S; that is, a spiraling outward toward S, for example, is ruled out. Furthermore, this property is uniform with respect to the initial position.

Returning now to the GSDS, to ensure that the whole funnel is repelled by S, it is reasonable to ask that condition (4) should hold for the worst possibility:

$$\sup_{t>0} \liminf_{\substack{y \to x \\ y \in X \setminus S}} \{\inf P(yt)/P(y)\} > 1 \qquad (x \in S) \tag{5}$$

Theorem 1

Assume that X is compact and let $S \subset X$ be compact with empty interior. Let $S, X \setminus S$ be forward invariant. Suppose that the continuous function $P: X \to \mathbb{R}^+$ is such that $P^{-1}(0) = S$. Then if condition (5) holds, there exists a compact absorbing set M for $X \setminus S$ with $d(S, M) > 0$.

The proof for an ordinary semidynamical system (Hutson, 1984) does not quite generalize as it stands, the difficulty coming from the possibility that for some subset X_0 the section xt may be neither wholly in X_0 nor in its complement X_0^c. The first of the preparatory lemmas below is a partial replacement for the idea that P should increase along orbits. In fact, it may first decrease, but must at some stage have increased with a certain uniformity with respect to position. Define

$$\alpha(t, x) = \begin{cases} [\inf P(xt)]/P(x) & (x \in X \setminus S) \\ \liminf\limits_{\substack{y \to x \\ y \in X \setminus S}} \alpha(t, y) & (x \in S) \end{cases}$$

This function is readily shown to be lower semicontinuous (LSC). Note also that from condition (5), for each $x \in S$, there is a t_x such that $\alpha(t_x, x) > 1$.

Lemma 2

There are a finite set $F = \{t_1, \ldots, t_k\} \subset (0, \infty)$, an open neighborhood U of S, and an $\bar{h} > 0$ such that the following holds. Given any $x \in U \setminus S$, there exists $t(x) \in F$ such that

$$\inf P\{\pi[x, t(x)]\} \geq (1 + \bar{h})P(x)$$

Proof

For $h > 0$, $t > 0$, the sets

$$U(h, t) = \{x : \alpha(t, x) > 1 + h\}$$

are open, since $\alpha(t, \cdot)$ is LSC. By the remark preceding *Lemma 2*

$$S \subset \bigcup_{h, t > 0} U(h, t)$$

and since S is compact, there exist $h_1, \ldots, h_k > 0$ and $t_1, \ldots, t_k$ such that

$$S \subset \bigcup_{i=1}^{i=k} U(h_i, t_i)$$

However, $U(a, t) \subset U(b, t)$ if $b < a$, so with $\bar{h} = \min h_i$, say,

$$S \subset \bigcup_{i=1}^{i=k} U(h, t_i) = U$$

Finally, if $x \in U \setminus S$, then $x \in U[\bar{h}, t(x)]$ for some $t(x) \in F$, so $\alpha[t(x), x] > 1 + \bar{h}$. This yields the result on using the definition of α.

Choose $p > 0$ such that

$$W_p \doteq \{x : 0 < P(x) \leq p\} \subset U$$

and set

$$N = X \setminus \bar{W}_p \qquad \bar{t} = \max_{t \in F} t$$

We note that $\bar{W}_p$ is a neighborhood of S, $N = \{x : P(x) > p\}$, and $\partial N \subset \{x : P(x) = p\}$.

Lemma 3

If $\gamma^+(\bar{N}) \setminus \bar{N} \neq \phi$, given any y in this set, there exist $z^* \in \partial N$, $t^* \in (0, \bar{t})$ such that:

(1) $y \in \pi(z^*, t^*)$.
(2) $\pi(z^*, t) \cap \bar{N}^0 \neq \phi$ for $0 < t < t^*$.

Proof

The idea is to show that there is a $z^* \in \bar{N}$ such that the time t^* from z^* to y is a minimum. Define

$$t^* = \inf\{t : y \in \pi(z, t) \text{ for some } t > 0,\ z \in \bar{N}\}$$

and note that from the assumption concerning y, this set is not empty. Thus, there are sequences (t_n) t^* and $(z_n) \in \bar{N}$ with $y \in \pi(z_n, t_n)$, and since $\bar{N}$ is compact, by selecting a subsequence if necessary, it follows that there is a $z^* \in \bar{N}$ such that $z_n \to z^*$. If $y \notin \pi(z^*, t^*)$, since $\pi(z^*, t^*)$ is compact, there is a neighborhood U of $\pi(z^*, t^*)$ such that $y \notin U$. But from the upper semicontinuity of π, there is a neighborhood V of $(z^*, t^{*)}$ such that $\pi(z, t) \subset U$ for $(z, t) \in V$. Since $(z_n, t_n^*) \in V$ for large enough n, $y \in \pi(z_n, t_n) \subset U$. This is a contradiction and shows that $y \in \pi(z^*, t^*)$.

If (2) of *Lemma 3* does not hold, there is a $t_1 \in (0, t^{*)}$ with $\pi(z^*, t_1) \subset \bar{N}$ and

$$\pi[\pi(z^*, t_1), t^* - t_1] = \pi(z^*, t^*)$$

It follows that there is a $z_1 \in \pi(z^*, t_1) \subset \bar{N}$ such that $y \in \pi(z_1, t^* - t_1)$. This contradicts the minimality of t^*, and (2) follows.

To prove that $z^* \in \partial N$, suppose on the contrary that $z^* \in N$ (open). Then, as $\pi(z^*,0) = z^*$ and π is USC, there is a neighborhood V of $(z^*,0)$ such that $\pi(z,t) \subset N$ for $(z,t) \in V$, and in particular $\pi(z^*,t_0) \subset N$ for some $t_0 \in (0,t^*)$. This contradicts (2) of *Lemma 3* and shows that $z^* \in \partial N$.

Since $z^* \in \partial N$, then $P(z^*) = p$ and there exists $t(z^*) \leq \bar{t}$ such that

$$\inf P\{\pi[z^*,t(z^*)]\} \geq (1 + \bar{h})p > p$$

whence $\pi[z^*,t(z^*)] \subset N$. From (2) of *Lemma 3* $t^* \leq t(z^*) \leq \bar{t}$.

Lemma 4

$\gamma^+(\bar{N})$ is a compact forward-invariant set with $d[S,\gamma^+(\bar{N})] > 0$.

Proof

From Aubin and Cellina (1984), $Y \doteq \pi(\bar{N},[0,\bar{t}])$ is compact. To prove that $\gamma^+(\bar{N}) = Y$, since clearly $Y \subset \gamma^+(\bar{N})$, it is enough to show that $\gamma^+(\bar{N}) \subset Y$. Take any $y \in \gamma^+(\bar{N})$, and note that if $y \in \bar{N}$, then $y \in Y$. If $y \in \gamma^+(\bar{N})\backslash\bar{N}$, by *Lemma 3* there are a $z^* \in \bar{N}$ and $t^* \in (0,\bar{t}]$ with $y \in \pi(z^*,t^*) \subset Y$, so $\gamma^+(\bar{N}) = Y$. Since $X\backslash S$ is forward invariant, $\gamma^+(\bar{N}) \cap S = \phi$, and the result follows from the compactness.

Lemma 5

Given $x_0 \in W_p$, there exists T such that $\pi(x_0,T) \subset \gamma^+(\bar{N})$.

Proof

Put $P(x_0) = p_0$, take n such that $(1 + \bar{h})^n p_0 > p$, and choose $T > n\bar{t}$. If $x_0 \in \bar{N}$ the result follows, so suppose $x_0 \in X\backslash N$ and take any $z \in \pi(x_0,T)$; it will be proved that $z \in \gamma^+(\bar{N})$.

With $t(\cdot)$, as in *Lemma 2*,

$$z \in \pi(x_0,T) = \pi\{\pi[x_0,t(x_0)], T - t(x_0)\}$$

Hence $z \in \pi[x_1,T - t(x_0)]$ for some $x_1 \in \pi[x_0,t(x_0)]$. If $x_1 \in \bar{N}$, again the result follows, otherwise the argument is repeated to yield a sequence $x_0,x_1,\ldots,x_j$. If $x_j \in \bar{N}$ for some $j < n$, the proof is complete. If not, since for each j, $nt(x_j) \leq n\bar{t} < T$, then

$$x_n \in \pi[x_0,t(x_0) + \cdots + t(x_{n-1})]$$

$$z \in \pi[(x_n,T - t(x_0) - \cdots - t(x_{n-1})]$$

and by *Lemma 2*,

$$P(x_n) \geq (1 + \bar{h})P(x_{n-1}) \geq \cdots \geq (1 + \bar{h})^n P(x_0) > p$$

Therefore, $x_n \in \bar{N}$ and $z \in \gamma^+(\bar{N})$, which proves *Lemma 3*.

To prove the theorem, note that by *Lemma 5*, $\gamma^+(\bar{N})$ is absorbing, and $d[S, \gamma^+(\bar{N})] > 0$ by *Lemma 4*.

4. Permanent Coexistence

It is now shown how *Theorem 1* may be applied to the differential inclusions (2) arising in a biological context. Take $G_i(x) = x_i F_i(x)$ and assume that G is USC with compact convex values. A solution is an absolutely continuous function that satisfies (2) and it is assumed that there is a nice compact neighborhood of the origin in $\mathbb{R}^+_n$ which is absorbing.

Take the GSDS to be composed of solutions $x(t)$. Let "dot" denote differentiation along a solution, and ∂_i partial differentiation. For P, a C^1 function,

$$P(xt)/P(x) = \exp\left\{ \int_0^t \dot{P}[x(s)]/P[x(s)]\, ds \right\}$$

$$= \exp\left\{ \int_0^t \sum_{i=1}^{i=n} \partial_i \log P[x(s)]\, \dot{x}_i(s)\, ds \right\}$$

Define the USC map $\Psi : X \setminus S \to A(\mathbb{R})$, and the LSC function $\psi : X \to \mathbb{R}$ by setting

$$\Psi(x) = \left\{ \sum_{i=1}^{i=n} \beta_i \partial_i P(x)/P(x) : \exists\ \beta_i \in G_i(x) \right\}$$

$$\psi(x) = \begin{cases} \inf \Psi(x) & (x \in X \setminus S) \\ \liminf\limits_{\substack{y \to x \\ y \in X \setminus S}} \psi(y) & (x \in S) \end{cases}$$

Lemma 6

Assume that:

(1) ψ is bounded below.
(2) $\psi(x) > 0 \qquad [x \in \overline{\Omega(S)}]$.

Then condition (5) of the theorem holds and the system (2) is permanently coexistent.

The important point here is that (2) of *Lemma 3* need only hold on $\overline{\Omega(S)}$. This is plausible and easy to prove, since for large enough t each semiorbit is near $\Omega(S)$ where $\psi > 0$, so $\exp(\psi) > 1$, yielding condition (5).

The following simple example is intended to illustrate, without involving complicated algebra, how the result may be applied to a specific system. Consider the following model of a predator–prey system, the f_i being perhaps errors in the experimental observations or climatic variations.

$$\dot{x}_1 = x_1[a - bx_1 - cx_2 + f_1(x,t)]$$

$$\dot{x}_2 = x_2[-c + dx_1 + f_2(x,t)]$$

where $a, b \cdots > 0$. Assume the following finite:

$$f_i = \inf_{x \in \mathbb{R}^+, t \geq 0} f_i(x,t)$$

$$\bar{f}_i = \sup_{x \in \mathbb{R}^+, t \geq 0} f_i(x,t)$$

and make the assumptions, natural for a predator–prey system, that

$$a + f_i > 0$$
$$-c + \bar{f}_2 < 0$$

Theorem 7

The system is permanently coexistent if

$$(a + f_1)/b > (c - f_2)/d$$

Proof

With $P(x) = x_1^{\alpha_1} x_2^{\alpha_2}$,

$$\Psi(x) = \{\alpha_1[a - bx_1 - cx_2 + \delta_1] + \alpha_2[-c + dx_1 + \delta_2]:$$

$$f_1 \le \delta_1 \le \bar{f}_1, f_2 \le \delta_2 \le \bar{f}_2\} .$$

Then

$$\Omega(S) \subset \{(0,0)\} \cup \{[(a + f_1)/b, (a + \bar{f}_1)/b], 0\}$$

It is a matter of simple algebra to show that under the assumed conditions there exist $\alpha_1, \alpha_2 > 0$ such that $\psi > 0$ on $\Omega(S)$. Thus permanent coexistence follows from *Lemma 6*.

The above result is not itself of great significance, but we remark that a similar analysis should be possible for much less tractable systems involving a larger number of species, e.g., two prey and a predator, although at the expense of considerable extra algebraic computation. The discussion shows that concrete results can be obtained for certain fundamental problems concerning systems which are only specified quite loosely.

Acknowledgment

The authors are grateful to Jean-Pierre Aubin for discussions concerning this problem.

References

Amann, E. and Hofbauer, J. (1984), Permanence in Lotka–Volterra and replicator equations, in M. Peshel (Ed), *Lotka–Volterra Approach to Dynamical Systems* (Akademie-Verlag, Berlin).

Aubin, J.P. and Cellina, A. (1984), *Differential Inclusions* (Springer-Verlag, Berlin).

Barbashin, E.A. (1948), On the theory of generalized dynamical systems, *Uchen. Zap. Moskov. Univ. no. 135, Mat.*, **2**, 110–133.

Butler, G., Freedman, H.I. and Waltham, P. (1985), *Uniformly Persistent Systems*, to appear in *Proc. Amer. Math. Soc.*

Dochev, D. (1979), Invariant sets in a general dynamical system, *Godischnik Vissh. Uchebn. Zaved. Prilozhna Mat.*, **14**, 179–187.

Hofbauer, J. (1981), General cooperation theorem for hypercycles, *Monatsh. Math.*, **91**, 233–240.

Hofbauer, J. (1985), Stability Concepts for Ecological Differential Equations, paper presented at a recent meeting.

Hutson, V. (1984), A theorem on average Ljapunov functions, *Monatsh. Math.*, **98**, 267–275.

Hutson, V. and Moran, W. (1982), Persistence of species obeying difference equations, *J. Math. Biol.*, **15**, 203–213.

Hutson, V. and Moran, W. (1985), Persistence in systems with diffusion, in K. Sigmund (Ed), *Dynamics of Macrosystems* (Akademie-Verlag, Berlin).

Kloeden, P.E. (1979), The funnel boundary of multivalued dynamical systems, *J. Australian Math. Soc.*, **27A**, 108–124.

Laricheva, G.A. (1984), Dynamical systems generated by quasihomogeneous multivalued mappings, *Soviet Math. Dokl.* **29**, 430–433.

Roxin, E. (1965), Stability in general control systems, *J. Differential Equations*, **1**, 115–150.

Schuster, P., Sigmund, K., and Wolff, R. (1979), Dynamical systems under constant organization. III Cooperative and competitive behaviour of hypercycles, *J. Differential Equations*, **32**, 357–368.

Szegö, G.P. and Treccani, G. (1969), *Semigruppi di Trasformazioni Multivoche*, Lecture Notes in Mathematics, Vol. **101** (Springer-Verlag, Berlin).

Stability for a Linear Functional Differential Equation with Infinite Delay

J. Milota
Charles University, Prague, ČSSR

1. Introduction

The Volterra model of a "historical action" in population growth,

$$\dot{u}(t) = au(t)[1 - bu(t) - c \int_0^{+\infty} k(s)u(t - s)ds] \tag{1}$$

yields a classical example of a functional differential equation with infinite delay. Sometimes it is necessary to add the Laplacian Δu to the right-hand side of equation (1) to express the effects of diffusion. Equation (1) then belongs to the class of semilinear parabolic equations,

$$\dot{u}(t) + Au(t) = F[u(t), u_t] \tag{2}$$

with infinite delay, i.e., u_t denotes the function on the interval $(-\infty, 0]$ given by $u_t(s) = u(t+s)$. The operator A stands here for $-\Delta u$ and therefore we suppose that A is a sectorial operator in some Banach space X (see, e.g., Friedman, 1969, or Henry, 1981). Then $-A$ is the infinitesimal generator of an analytic semigroup e^{-At} and the spectrum of A is in a sector $\{\lambda \in C, \mid \arg(\lambda - \alpha) \mid < \vartheta\}$, where $\vartheta < \pi/2$. The vertex α will be specified later on.

The right-hand side F in the Volterra model depends only on values of u and does not depend on its spatial derivatives. But there are very important examples of equation (2) in which F depends even on the highest spatial derivatives of u, e.g., equations for heat conduction in materials with memory. This dependence can be specified with the help of fractional powers A^α. We denote by X^α the domain of A^α endowed with the graph norm. Further, Y^α denotes a space of functions defined on the interval $(-\infty, 0]$ with values in X^α. Function spaces Y^α will be equipped with norms of a fading memory type (see, e.g., Coleman and Mizel, 1966). These norms have fundamental importance for the stability theory for equations with infinite delays, as has been shown by Hale and Kato (1978), Schumacher (1978), and, more recently, Kappel and Schappacher (1980).

We suppose that the nonlinear term F in equation (2) maps some subset of $X^\alpha \times Y^\beta$ into X. Roughly speaking, the cases $0 < \alpha < 1$ or $0 < \beta < 1$ correspond to a dependence of F on lower spatial derivatives of u. If F depends on the highest spatial derivatives, then it is necessary to take $\alpha = 1$ or $\beta = 1$. These cases cause difficulties in proving the existence and continuous dependence results for the Cauchy problem regarding lack of uniqueness of a solution (see Milota and Petzeltová, 1985a, b, and the papers cited there). Equation (2) generates a C_0-semigroup only under special assumptions on F even for a linear map F and a finite delay (see Kunish and Schappacher, 1983, and di Blasio *et al.*, 1984).

It is our purpose here to investigate properties of the solution operator for equation (2) with linear F. Such properties yield information on solutions of a nonlinear equation near to an equilibrium point (linearized stability etc.). We therefore consider the Cauchy problem given by

$$\begin{aligned} &\dot{u}(t) + Au(t) = Lu_t \qquad t > 0 \\ &u_0 = \varphi \end{aligned} \tag{3}$$

Here A is a sectorial operator with $\alpha > 0$. This assumption means that, for $A = -\Delta$, generally only Dirichlet boundary conditions are allowed. Moreover, L is supposed to be a linear continuous operator of Y^α into X. We should remark that our method requires the restrictive assumption $\alpha < 1$. As regards the spaces Y^α, we take some $\gamma > 0$ and define Y^α to be the space of all continuous maps φ of the interval $(-\infty,0]$ into X^α for which the norm

$$|\varphi|_{Y^\alpha} := \sup_{s \le 0} |e^{\gamma s}\, \varphi(s)|_{X^\alpha}$$

is finite. For further purposes we denote all these assumptions by (H1).

2. A Solution Semigroup

A continuous function $u : (-\infty,T) \to X^\alpha$ is called a mild solution to equation (3) if $T > 0$ and u satisfies:

$$\begin{aligned} &(1)\quad u_t \in Y^\alpha \qquad \text{for each} \qquad t \in (0,T). \\ &(2)\quad u_0 = \varphi. \\ &(3)\quad u(t) = e^{-At}\, \varphi(0) + \int_0^t e^{-A(t-s)} Lu_s \, ds \\ &\qquad \text{for each } t \qquad \in (0,T). \end{aligned} \tag{4}$$

Proposition 1

Let (H1) be satisfied. Then for any $\varphi \in Y^\alpha$ there exists a unique mild solution of equation (3). This solution is defined on the interval $(-\infty, +\infty)$.

The proof is rather standard. The contraction principle ($\alpha < 1$) yields a local solution. The uniqueness and the global existence follow from an estimate of the Gronwall type. See Milota (1986) for more details of this and further proofs.

Corollary

Let (H1) be satisfied, $\varphi \in Y^\alpha$, and let $u(.,\varphi)$ denote a mild solution of equation (3) on the interval $(-\infty,+\infty)$. If $T(t)\varphi$ denotes $u_t(\varphi)$, then $T(t)$ is a C_0-semigroup on the space Y^α.

We note that the solution semigroup $T(t)$ is not generally an analytic semigroup. To examine the asymptotic behavior of $T(t)$ we introduce the pure parabolic problem

$$\dot{v}(t) + Av(t) = 0 \qquad t > 0$$
$$v_0 = \varphi$$

and denote by $S(t)$ its solution semigroup in the space Y^α.

Proposition 2

Let (H1) be satisfied together with the hypothesis (H2) that A has a compact resolvent in the space X. Then the operator $T(t) - S(t)$ is a compact map of Y^α into Y^α for each t positive.

The proof is based on the Arzèla–Ascoli theorem. The assumption (H2) implies that the embedding of X^β into X^α is compact for $\beta > \alpha$. As the integral operator in equation (4) maps bounded sets in Y^α into bounded sets in $X^{\alpha+\varepsilon}$ ($\alpha < 1$), the pointwise compactness follows.

3. The Spectrum of $T(t)$

In order to estimate the norm of $T(t)$ one commonly has to characterize the infinitesimal generator B of $T(t)$ and its spectrum. But the main difficulties with infinite delays consist in the fact that this generator is not completely described even for $X = R^n$. In this case, Naito (1979) showed how to overcome these obstacles. It is sufficient to estimate the radius of the essential spectrum of $T(t)$ and to calculate the point spectrum $P_\sigma(B)$ of the generator B. The notion of the essential spectrum is used in the sense of Browder (1961). Because of the following two lemmas we can use the same procedure also for an infinitely dimensional space X.

Lemma 1

Under the hypotheses (H1) and (H2), the estimate

$$r_e[T(t)] \leq \text{const.} e^{-\min(\alpha,\gamma)t}$$

holds for the radius of the essential spectrum of $T(t)$.

The proof follows easily from the Nussbaum (1970) formula for the radius r_e and *Proposition 2*.

The second lemma is more technical.

Lemma 2

Let (H1) be satisfied. Then $\lambda \in P_\sigma(B)$ if and only if $\operatorname{Re} \lambda \geq -\gamma$ and the characteristic equation

$$\lambda x + Ax - L(e^{\lambda s} x) = 0 \tag{5}$$

has a nontrivial solution in $D(A)$.

With respect to the estimate of the resolvent of a sectorial operator one can deduce from *Lemma 2* that the set $P_\sigma(B) \cap \rho(A)$ is bounded. This fact together with *Lemma 1* yields the following conclusion.

Lemma 3

Let the hypotheses (H1) and (H2) be satisfied. Then for any $\varepsilon > 0$ the set $\{\lambda \in C,\ \operatorname{Re} \lambda > -\min(\alpha,\gamma) + \varepsilon\}$ contains a finite number of points of $P_\sigma(B)$ only, and all these points are of finite multiplicity.

Proposition 3 (asymptotic stability)

Let the hypotheses (H1) and (H2) be satisfied and let $\operatorname{Re} \lambda < 0$ whenever the characteristic equation (5) has a nontrivial solution. Then there is a $\delta > 0$ and a constant c such that

$$\| T(t) \| \leq c e^{-\delta t}$$

for all $t > 0$.

The proof is based on an estimate of the spectral radius of $T(t)$ which can be derived from *Lemma 1* and the well-known relation between the point spectra of a C_0-semigroup and its generator.

It is possible to obtain a more precise result than *Lemma 2*, namely the generalized eigenspaces of B can be characterized. This characterization implies that there are projectors on the generalized eigenspaces which commute with the solution semigroup $T(t)$. This is the main step in proving the saddle-point property of the zero solution of equation (3).

Proposition 4

Let the assumptions (H1) and (H2) be satisfied. Then there exists a decomposition $Y^a = Y_1 \oplus Y_2$ such that:

(1) Y_1 has a finite dimension.
(2) Y_1, Y_2 are $T(t)$-invariant.
(3) The zero solution is asymptotically stable for $T(t)_{/Y_2}$.
(4) $Y_1 \subset D(B)$ and $B_{/Y_1}$ is a continuous linear operator that generates a group which is an extension of $T(t)_{/Y_1}$.

References

Blasio, G. di, Kunish, K., and Sinastrari, E. (1984), *Stability for Abstract Linear Functional Differential Equations*, preprint (Universität Graz, Graz, Austria).

Browder, F. E. (1961), On the spectral theory of elliptic differential operators I, *Math. Ann.*, **142**, 22–130.

Coleman, B. D. and Mizel, V. J. (1966), Norms and semigroups in the theory of fading memory, *Arch. Rat. Mech. Anal.*, **23**, 87–123.

Friedman, A. (1969), *Partial Differential Equations* (Holt, Rinehart, and Winston, New York).

Hale, J. K. and Kato, J. (1978), Phase space for retarded equations with infinite delay, *Funkcial. Ekvac.*, **21**, 11–41.

Henry, D. (1981). Geometric Theory of Semilinear Parabolic Equations, *Lecture Notes in Mathematics, No. 840* (Springer, Berlin).

Kappel, F. and Schappacher, W. (1980), Some considerations to the fundamental theory of infinite delay equations, *J. Diff. Eqs.*, **37**, 141–183.

Kunish, K. and Schappacher, W. (1983), Necessary conditions for partial differential equations with delay to generate C_0-semigroup, *J. Diff. Eqs.*, **50**, 49–79.

Milota, J. (1986), Stability and saddle-point property for a linear autonomous functional parabolic equation, *Comm. Math. Univ. Carolinae*, **27** (to appear).

Milota, J. and Petzeltová, H. (1985a), An existence for semilinear functional parabolic equations, *Cas. Pěst. Mat.*, **110**, 274–288.

Milota, J. and Petzeltová, H. (1985b), Continuous dependence for semilinear parabolic functional equations without uniqueness, *Cas. Pěst., Mat.*, **110** (to appear).

Naito, T. (1979), On linear autonomous retarded equations with an abstract phase space for infinite delay, *J. Diff. Eqs.*, **33**, 74–91.

Nussbaum, R. (1970), The radius of the essential spectrum, *Duke Math. J.*, **37**, 473–478.

Schumacher, K. (1978), Existence and continuous dependence for functional-differential equation with unbounded delay, *Arch. Rat. Mech. Anal.*, **67**, 315–335.

III. STABILITY ANALYSIS

The Ljapunov Vector Function Method in the Analysis of Stability and other Dynamic Properties of Nonlinear Systems

V.M. Matrosov
Irkutsk Computer Center of the Siberian Division of the USSR Academy of Sciences

A method of analysis of some dynamic properties of nonlinear systems of various types developed at the Irkutsk Computer Center is given. Applications of this method showed that it has high efficiency, considerable advantages over other known methods of stability analysis of nonlinear systems, and a need to be expanded into a more general form. It has been possible to extend this method to abstract concepts of dynamics and control theory. The problem of deriving theorems from the vector of the Ljapunov function method for the dynamic properties of various types is discussed. Methods and algorithms for the construction and application of the Ljapunov vector function (LVF) are described.

1. Abstract Concepts of Dynamics

Concepts that involve known dynamic models of different types (Matrosov and Anapolski) are described. The main variables of abstract dynamics are introduced: $t \in T$ (current time), $t_0 \in T_0$ (initial time), $\mathbf{x} \in \mathbf{X}^t$ (current state), $h_{t_0} \in H_{t_0}$ (input), $h = (t_0, h_{t_0}) \in H$ (initial data), $\mathbf{x} : (T) \to \bigcup_{t \in T} \mathbf{X}^t$ (partial function for processes), $(t, x) \in \Xi$ (position). The sets are:

T – partially ordered set by $\leq$, $T_0 \subseteq T$.
$\mathbf{X}^t$ – state space.
H_{t_0} – input space.
$\Xi \triangleq \{(t, \mathbf{x}) : t \in T, \mathbf{x} \in \mathbf{X}^t\}$ – positions state.
$\Phi \triangleq \{x : x(t) \in \mathbf{X}^t, \operatorname{dom} x \subseteq T\}$ – space of processes.
$H \triangleq \{(t_0, h_{t_0}) : t_0 \in T_0, h_{t_0} \in H_{t_0}\}$ – initial data space.

The relation $r \subset \Phi \times H$ is called the system of processes (SP), if the following initial data axiom hold:

$(\forall h \in \operatorname{dom} r \ \forall x \in rh)\ t_0 \in \operatorname{dom} x \Lambda\{x(t_0)\}$ – single.

$x \in h$ is called the process of SPr with initial data h,

$$F(h,t) \triangleq \{\mathbf{x} \in \mathbf{X} : (\exists x \in rh)\mathbf{x} = x(t)\}, \qquad \mathbf{X} = \bigcup_{t \in T} \mathbf{X}^t$$

The concept of SP involves classical dynamic systems in metric or topological space (Markov and Nemitskii), general dynamic systems (Barbashin), general systems (nonautonomous) (Zubov), semidynamic systems, abstract processes (Haek), abstract numerical processes (Babushka, Prager, and Vitacek), polysystems (Bushan), sets of solutions of the differential, integral, integro-differential, difference-differential, and difference equations, stochastic differential equations, difference schemes, and so on.

SPr is called the abstract controlled system (ACS), if $h_{t_0} = (h^0_{t_0}, u, p)$, where $h^0_{t_0} \in H^0$ (initial states), $u \in U$ (admissible controls), $p \in P$ (perturbations); U, P are some functional spaces. The set of solutions of classical control systems, the abstract dynamic controlled system (Kalman), and so on, are involved.

The behavior of the processes $x \in rh$ is stated with respect to the nonmain variables : $P \in \mathbf{R}_q \subset 2^{\Xi}$, $P^0 \in \mathbf{R}^0_q \subset 2^H$ (current and initial estimation sets) $(q = 1, 2)$.

Our unified representation of the definitions B of dynamic properties is basically the following: they are formed from the same formula $\mathbf{R}_q$ $(q = 1, 2)$ using the connectives $\wedge, \vee$ and typed quantifiers $\mathbf{W}^{\mu} : \hat{\mathbf{W}}^{\mu} \triangleq (\forall z^{\mu} : Z^{\mu})$, $\overset{\vee^{\mu}}{\mathbf{W}} \triangleq (\exists z^{\mu} \in Z^{\mu})$. For example, let

$$R^l_+ \triangleq \{\varepsilon \in R^l : \varepsilon > 0\} \qquad \rho : \Xi \to R^l_+ \ \rho^0 : H \to R^{l0}_+$$

$\mathbf{B}_+$ – class of positive $l \times l_0$ matrices.

$\alpha \in R^1_+$

$B \in \mathbf{B}_+$

$\gamma \in R^l_+$

$\delta \in R^{l0}_+$

$T = R^1_+$

$P^0_{t0}(\delta) \triangleq \{h_{t_0} \in H_{t_0} : \rho^0(t_0, h_{t_0}) < \delta\}$

$T_{t_0} \triangleq \{t \in T : t_0 \le t\}$

$P^t(\alpha, B, \gamma, h) \triangleq \{\mathbf{x} \in \mathbf{X}^t : \rho(t, \mathbf{x}) \le \gamma + B\rho^0(h) \exp[-\alpha(t - t_0)]\}$

The definition of the exponential invariance of r for fixed $\gamma \in R^l_+$ is the following:

$$\left[\forall t_0 \in T_0\ \exists \alpha \in R_+^1\ \exists B \in \mathbf{B}_+\ \exists \delta \in R_+^{l0}\right.$$
$$\left.\forall h_{t_0} \in P_{t_0}^0(\delta)\ \forall x \in rh\ \forall t \in T(x)\ \forall \mathbf{x} = x(t)\right]$$
$$\mathbf{x} \in P^t(\alpha, B, \gamma, h)$$

Here $T(x) \triangleq T_{t_0} \cap \operatorname{dom} x$. If $\gamma = 0 \in R^l$ we have the exponential stability.

Let γ be not fixed for ACSr, but there exist $u \in U$, $\gamma \in R_+^l$ such that for any $p \in P$ the formulae are executed. Then we have an exponential boundedness of ACSr (Matrosov *et al.*, 1980).

2. Comparison Systems and Vector Functions

Let us introduce the following auxiliary sets and variables: $h_c \in H_c$, $\mathbf{x}_c \in \mathrm{X}_c^t$, $x_c \in \Phi_c$,... and the functions $v_q : (t, x) \to v_q(t, x)$, $\Xi \to X_0$ $(q = 1, 2)$, where X_c^t is partially ordered by $\preceq$. The SP $r_c \subset \varphi_c \times H_c$ and function v are called the comparison system (CS) and vector comparison (VCF) for SPr, respectively, if the inequalities are satisfied.

$$W \triangleq \Lambda \{\forall t_0 \in T^0\ \forall h_{t_0}\ \ \forall h_{t_{0c}} \in H_{t_{0c}} \in H_{t_{0c}} : x_{0c} = v_q[t_0, x(t_0)]\}$$
$$\forall x \in rh \qquad \exists x_c \in r_c h_c \qquad \forall t \in T(x) \cap T_c(x_c) v_q[t, x(t)] \preceq x_c(t)$$
$$(q = 1, 2)$$

(the connection formulae).

In actual cases we can obtain the sufficient conditions for these using theorems on differential, integral, finite-difference, and operator inequalities (Matrosov, 1973).

Let, for instance, SPr be the set of the Carateodory type solutions of the differential equation

$$\frac{dx}{dt} = X(t, x) \tag{1}$$

in Banach space E with the unbounded discontinuous operator $\mathrm{X} : \Omega \to E$, $\Omega = T \times H_0$, $T = [0, +\infty)$, $H_0 \subseteq E$. Let the function $v : \Omega \to R^k$ satisfy the differential inequality for the solutions of equation (1)

$$\bar{D}^+ v[t, x(t)] \preceq f\{t, v[t, x(t)]\} \tag{2}$$
$$t \in T(x)$$

with a function $f : A \to R^k$ that satisfies the conditions of "quasi-monotony":

$$f^s(t,y_1) \le f^s(t,y_2)$$

for

$$y_1^s = y_2^s$$
$$y_1^\nu \le y_2^\nu \qquad (\nu \ne s) \qquad (s,\nu = \overline{1,k}) \qquad (A \le T \times R^k)$$

of the boundedness and absolute upper semicontinuity of $v[\cdot,x(\cdot)]$ on any compact space. Then we can prove that the set of the upper solutions x_c of the equation

$$\frac{dy}{dt} = f(t,y) \qquad y \in R^k \tag{3}$$

can understand the solutions of equation (3) in the sense of OII-solutions, which satisfy the differential inequalities (Matrosov, 1971)

$$\lim_{y' \to y(t)} f(t,y') \le \frac{dy(t)}{dt} \le \overline{\lim_{y' \to y(t)}} f(t,y') \qquad t \dot{\in} T(y)$$

For every dynamic property B of SPr an auxiliary property B_c of CS is introduced. For example, we can use the exponential upper semiinvariance of CS (for fixed $\gamma_c \in R_+^k$)

$$\Big[\forall t_0 \in T_0\ \exists \alpha_c \in R_+^1\ \exists \beta_c \in R_+^k\ \exists \delta_c \in R_+^k$$
$$\forall h_{t_0c} \in P^0_{t_0c}\ \ \forall x_c \in r_c h_c\ \ \forall t \in T_c(x_c)\Big]$$
$$x_c(t) \le \gamma_c + \beta_c \| h_{t_0c} \| \exp[-\alpha_c(t-t_0)]$$

We assume here that $H_{t_0c} = R^k$, $P^0_{t_0c}(\delta_c) \triangleq \{h_{t_0c} \in R^k : h_{t_0c} < \delta_c\}$. The same holds true for ACS.

3. The Comparison Principle

Theorems on stability of various types (Matrosov, 1962), on boundedness, dissipations, attraction, uniqueness of the processes, and many others, have been proved using VCF and CS. It is possible to combine the results of these theorems in the form of the principal idea of the comparison principle: if VCF and CS exist and satisfy the suitable conditions, then the various dynamic properties of nonlinear systems in question follow from the corresponding properties of CS (Matrosov, 1971).

As a result of an analysis of the formulations and proofs of hundreds of theorems of the Ljapunov function method and the comparison method (in the simplest case by Abets and Pfeiffer) connections between formulae of the properties studied and the conditions on LVF in the comparison theorems were revealed; a series of heuristics were denoted, of which the principal one

$$\mathrm{T}_q^{z\mu} \triangleq \begin{cases} \hat{\mathrm{W}}_q^{\mu}\ \overset{\vee}{\mathrm{W}}{}_{qc}^{\mu} & | \ \mathrm{W}_q^{\mu} = \hat{\mathrm{W}}_q^{\mu} \\ \hat{\mathrm{W}}_{qc}^{\mu}\ \overset{\vee}{\mathrm{W}}{}_{q\mu} & | \ \mathrm{W}_q^{\mu} = \overset{\vee}{\mathrm{W}}{}_q^{\mu} \end{cases} \qquad (q = 0,1,2)$$

is the transformation of typed quantifiers W_q^{μ} into the subformulae of requirements on LVF. An algorithmic comparison principle with LVF was proved for SP (Matrosov, 1973). Let us denote

$$\hat{x}_{0q}\ \hat{x}_{0qc}^{v} \triangleq [\forall \mathbf{x}_0 \in P_{t_0q} \cap v(t_0,\cdot)^{-1} P_{t_0qc}^0\ \forall \mathbf{x}_0 = v(t_0,\mathbf{x}_0)]$$

$$\hat{\mathrm{F}}_{*q} \triangleq (\forall t_{*q} \in T_{*q} \cap T_{*qc})\mathrm{F}_{t*q} \triangleq (\forall t \in T_{t*q} \cap T_{t*qc}) \qquad (q = 0, 1, 2)$$

$$\mathrm{T}_{t_0t_*}^{\varepsilon} \triangleq \begin{cases} \hat{\mathrm{F}}_0\ \hat{\mathrm{F}}_*\ \mathrm{T}^{\varepsilon} & | \ t_0 < t_* < \varepsilon \\ \hat{\mathrm{F}}_0\ \mathrm{T}^{\varepsilon}\ \hat{\mathrm{F}}_* & | \ t_0 < \varepsilon < t_* \\ \mathrm{T}^{\varepsilon}\ \hat{\mathrm{F}}_0\ \hat{\mathrm{F}}_* & | \ \varepsilon < t_0 < t_* \end{cases}$$

Here and later $\hat{\mathrm{F}}_0$ ($\hat{\mathrm{F}}_*$ or T^{ε}, ... respectively) is replaced by the empty formula Λ, if $t_0 \not\in \mathrm{B}$ (t_* or $\varepsilon \not\in \mathrm{B}$). $\mathrm{T}_{t*qt}^{\varepsilon q}$ is determined analogously.

The following formulae are introduced ($q = 1, 2$):

$$\mathrm{B}_q \triangleq \mathrm{T}_0\ \mathrm{T}_q\ \mathrm{R}_q$$

$$\mathrm{T}_{tq} \triangleq \hat{\mathrm{F}}_0 \begin{cases} \mathrm{T}^{t_*}\ \mathrm{T}^{t_{*q}} & t_*, t_{*Q} \in \mathbf{B}_q \\ \mathrm{T}^{t_*} & | \ t_* \in \mathbf{B}_q \\ \mathrm{T}_q^{t_{*q}} & | \ t_* \in \mathbf{B}_q, t_{*Q} \not\in \mathbf{B}_q \\ & | \ t_{*Q} \in \mathbf{B}_q, t_* \not\in \mathbf{B}_q \end{cases}$$

$$\sigma \begin{cases} P_{t_0q}^0, P_{t_0qc}^0 & \hat{x}_{0q}\hat{x}_{0qc}^{v} \mid \delta_{qc} V \delta \in \mathbf{B}_q \\ UP_{t_0q}^0\ UP_{t_0qc}^0 & \hat{x}_{0q}\hat{x}_{0qc}^{v} \mid \delta_q, \delta \in \mathbf{B}_q, x_0 \in \mathbf{B}_q \end{cases}$$

$$\begin{cases} \mathrm{T}_q^{x} & | \ x_q = \hat{x} \\ \hat{x}\overset{\vee}{x}_c & | \ x_q = \overset{\vee}{x} \end{cases} \qquad \begin{cases} T_q(x) \subseteq T_c(x_c) & | \ \mathrm{F}_q = \hat{\mathrm{F}} \\ T_c(x_c) \subseteq T_q(x) & | \ \mathrm{F}_q = \overset{\vee}{\mathrm{F}} \end{cases}$$

$$T_{\mathbf{x}q} \triangleq T^{\varepsilon}_{t_0 t},\ T^{\varepsilon_q}_{t_{*q} t}\left\{\begin{array}{ll}(\forall \mathbf{x}\in \hat{X}\, P^t_q\ \ \forall \mathbf{x}_c \in P^t_{qc}) v(t,\mathbf{x}) \le \mathbf{x}_c & \mid R_q \neq \Lambda \\ \Lambda & \mid R_q = \Lambda\end{array}\right\}$$

$$T_{\mathbf{x}_0 q} \triangleq \left\{\begin{array}{ll}T^{\delta}\,\hat{F}_0 & \mid \delta \prec t_0 \\ \hat{F}\,T^{\delta} & \mid t_0 \prec \delta\end{array}\right\}\quad\left\{\begin{array}{ll}T^{\delta_q} & \mid \delta_q \in B_q \\ \Lambda & \mid \delta_q \notin B_q\end{array}\right\}$$

$$\left\{\begin{array}{ll}v[t_0, P^*_q(\delta_q,\delta,t,t_0)] \subseteq P^*_{qc}(\delta_{qc},\delta_c,t_c) & \mid x_{0q} = \hat{x}_0 \\ P^*_{qc}(\delta_{qc},\delta_c,t_0) \subseteq v[t_0, P^*_q(\delta_q,\delta,t_0)] & \mid x_{0q} = x_0 \\ v(t_0,\mathbf{x}_0) = \mathbf{x}_c & \mid x_0 = B_q\end{array}\right\}$$

A VCF that satisfies (with CS) the conditions T_{tq}, T_{xxq}, $T_{\mathbf{x}_0 q}$ $(q = 1, 2)$ is called the LVF for equation (1) with respect to B.

3.1. Comparison principle

Let LVF v for SP with respect to B and CS r_c exist. Then the existence of dynamic property B in system (1) follows from the existence of the corresponding property B_c in CS r_c.

This explicit algorithmic form of the comparison principle in abstract dynamics gives the explicit scheme of the comparison theorems:

$$(\exists v, f) T_{\mathbf{x}q},\ T_{tq},\ T_{x_0 q} \qquad (q = 1, 2) \mapsto B_c \Longrightarrow B$$

A complete basis exists for establishing the comparison principle in abstract dynamics, abstract control theory, and mathematical control theory in the general form. It determines an algorithm for derivation of comparison theorems (Matrosov *et al.*, 1980). Such a theorem was obtained earlier only through a creative approach. The derivation of comparison theorems on the basis of the comparison principle is realized algorithmically by computer (Matrosov and Vassiliev, 1978).

Let, for example, $H^0_{t_{0c}} = H^0_c \subseteq X_c = X^{tc}_c$, $T_{0c} = T^0 \subseteq T = T_c$, and there are fixed $\gamma \in R_+$, $\gamma_c \in R^k_+$.

3.2. Comparison theorem

Let there exist VCF $v : \Xi \to X_c$ and CS r_c satisfying the conditions

$$\left[\forall t_0 \in T^0_0\ \ \forall \delta_c \in R^k_+\ \ \exists \delta \in R^{l0}_+\ \ \forall h_{t0} \in P^0_{t_0}(\delta)\right]$$

$$\times v[t_0 x(t_0)] \subseteq P^0_{t_{0c}}(\delta_c)$$

[the condition of upper semicontinuity of $v(t_0,\cdot)$ with respect to the families of sets $\{P^0_{t_0}(\delta)\}$, $\{P^0_{t_{0c}}(\delta_c)\}$],

$$\forall t_0 \in T^0 \qquad \forall h_{t_0} \in \cup \{P^0_{t_0}(\delta) : \delta \in R^{l0}_+\}$$
$$\forall x \in rh \qquad \forall h_{t_{0c}} = v[t_0, x(t_0) \qquad \forall x_c \in r_c h_c]\ T(x) \subseteq T_c(x_c)$$

(the condition of *continuability* of x_c to the right),

$$\Big[\forall t_0 \in T^0\ \forall \alpha_c \in R^1_+\ \exists \alpha \in R^1_+\ \ \forall \beta_c \in R^k_+\ \ \exists B \in \mathrm{B}_+$$
$$\forall h_{t0} \in \cup\{P^0_{t0}(\delta) : \delta \in R^{l0}_+\}\ \forall t \in T_{t0}\ \forall \mathbf{x} \in \mathbf{X}^t P^t(\alpha, B, \gamma, h)\Big]$$
$$v(t,x)! \le \gamma_c + \beta_c \| v[t_0, x(t_0)] \exp[-\alpha_c (t - t_0)]$$

(the main condition on function v of lower boundedness type), and CS r_c is exponentially upper semiinvariant, then SPr is exponentially invariant.

If $\gamma = 0$, $\gamma_c = 0$ in the main condition and CS r_c is upper exponentially semi-stable, then SPr is exponentially stable.

4. Derivation of the Theorems on Dynamic Properties (with LVF)

With the help of comparison theorems, dynamic properties analysis can be reduced to the LVF, CS construction and to essentially a simpler analysis of the corresponding properties of CS. The latter was realized for a sufficiently general case by Kozlov and Martinyuk *et al.* separately. A more effective theorem on the dynamic property B of system (1) was obtained. Some algorithms for a transition from comparison theorems to the theorem on dynamic properties (with LVF) are given by Matrosov (1981). More effective theorems can be obtained by adding some conditions $\mathrm{F}(\mathrm{B}_c)$, which are sufficient for the existence of property B_c in CS, to the comparison theorem and replacing W, D,... by simpler conditions W^*, D^*, F^*_δ ($j = 1, m+1, \delta = 1, \eta$) ($\mathrm{W}^*$ is expressed by differential, integral, difference, or operator inequalities).

4.2. The structure of the theorems on dynamic properties with LVF

Given LVF with respect to the dynamic property B for the SPr, then

$$\mathrm{F}(\mathrm{B}_c), \mathrm{W}^*, \mathrm{D}^*_1, \ldots, \mathrm{D}^*_{m+1}, \mathrm{T}^*_1, \ldots, \mathrm{T}^*_\eta \Rightarrow \mathrm{B}$$

The following theorem on exponential invariance (with given $\gamma \in R^l_+$) for differential equation (1) with differential inequality (2) and CS (3) is proved.

Theorem

Let LVF $v : T \times E \to R^k$ exist such that

$$\mu \cdot \max_{i=\overline{1,l}} [\rho^i(t,x)]^2 \le \max_{s=\overline{1,k}} v^s(t,x)$$

$$\| v[t_0, x(t_0)] \| \le \mu \cdot \| \rho^0(h) \|^2$$

$$v'(t,x) \le P[v(t,x) - \gamma_c] + \tilde{f}[t, v(t,x) - \gamma_c] \qquad (t,x) \in \Omega$$

$$p_{sj} \ge 0 \qquad (s \ne j)\ (-1)^s \cdot |p_{ij}|_1^s > 0 \qquad (s = \overline{1,k})$$

$$\tilde{f}(t, x_c - \gamma_c) / \| x_c - \gamma_c \| \underset{t \in T}{\rightrightarrows} 0 \qquad \text{if} \qquad x_c \to \gamma_c$$

$$\tilde{f}^s(t, x_c' - \gamma_c) \le \tilde{f}^s(t, x_c - \gamma_c)$$

for

$$x_c'^s = x_c^s \qquad x_c'^j \le x_c^j \qquad (j \ne s)\ (s = \overline{1,k})$$

$\tilde{f}$ satisfies the Carateodory condition and is bounded in every cylinder $T \times 0\eta$.

$$0 < \underline{\mu} < \bar{\mu}, \qquad \max_{s=\overline{1,k}} \gamma_c^s \le \underline{\mu} \left[\min_{i=\overline{1,l}} \gamma^i \right]^2$$

Then system (1) is exponentially invariant. If, in addition, $\gamma = 0$, $\gamma_c = 0$, then system (1) is exponentially stable.

A package of programs for the LVF theorem derivation method was elaborated on the basis of developed algorithms. More than 200 theorems on the LVF method for various dynamic properties of SP, ACS, and differential systems were obtained by computer at different stages of the work. Theorems obtained in such a way turned out to be either new theorems or modifications and generalizations of known ones (if prototypes existed). The latter corresponded fully to the theorems obtained manually. Good results were also obtained in the investigation of comparison theorems. Vasiliev (1979) developed an algorithm for the derivation of theorems on dynamic properties with necessary and sufficient conditions for dynamic properties formulae that include universal quantifiers on t –[$\forall t \in T(x)$].

Research in a new field of science, i.e., artificial intelligence, was initiated (Matrosov *et al.*, 1981).

5. The Construction of the Ljapunov Vector Function — Algorithms and Applications

We now describe three groups of methods elaborated for CS and VLF construction and some applications. We consider only problems of exponential stability, invariance, and boundedness for finite-dimensional, difference, and difference-differential equations (nonlinear).

In the first group of methods for CS and LVF v construction we use the Ljapunov–Poincaré lemma for autonomous systems

$$\dot{x} = X(x) = Bx + X(x) \qquad x \in R^n \tag{4}$$

with a holomorphic right-hand side in nonresonance cases

$$\operatorname{Re} \lambda_s(B) < 0$$

$$\lambda_s \neq \sum_{i=1}^{n} m_i \lambda_i \qquad m_i = 0,1,2,\ldots, \sum_{i=1}^{n} m_i > 1 \qquad (s = \overline{1,n})$$

CS is described by the equation

$$\dot{v}^s(x) = \kappa_s v^s(x) \qquad (\kappa_s = \text{const}\,,\ \kappa_s < 0\,,\ \tilde{f} \equiv 0)$$

The holomorphic LVF $v(x) = [v^1(x), \ldots, v^k(x)]$, $v^s(x) = z^s(x)\bar{z}^s(x)$ is constructed so as to satisfy the precise exponential estimations

$$v^s[x(t)] = v^s(x_0) \exp[\kappa_s(t - t_0)]$$

and is based on a solution of the equations with partial derivatives

$$\nabla z^s(x) \cdot X(x) = \lambda_s z^s(x)$$

The existence of holomorphic solutions is given by the Ljapunov–Poincaré lemma.

Such an LVF determines the isochrones of the system that give a precise description of the evolution of some neighborhood. An algorithm of quadratic LVF construction for the first approximation of differential equation (4) was implemented.

Methods of LVF construction to satisfy nonlinear differential equations of comparison are based on results by Persidski, Waleev, and Finin. The principal feature of this group of methods is the high precision of estimations obtained by LVF, but difficulties did exist in their applications due to the complexity of the algorithm's realization by computer. More recently, Kozlov

et al. have developed a more efficient algorithm for the construction of CS and LVF with components of a linear form module $v^s = |V^s \cdot x|$ for nonlinear systems and estimations

$$v(\mathbf{x}) \le |T|^{\vee} \cdot |\mathbf{x}| \qquad |\mathbf{x}| \le |s|^{\wedge} \cdot v(\mathbf{x})$$

An advantage of this method is the possibility of CS construction in an explicit form

$$\frac{dx_c}{dt} = p \cdot x_c + k \cdot \max\{Mx_c - u, |c| \cdot \max(Cx_c - \bar{u})\} + R$$

even for discontinuous nonlinearities. Right-hand sides are usually piecewise linear or polynomial [for polynomial or holomorphic $X(x)$].

For linear systems precise exponential estimations for this method exist. Algorithms were developed for the analysis of stability of interconnected control systems under perturbations. These algorithms were applied to electroenergetic systems (EES) and studies on gyroscopic stabilizers and orbital radiotelescopes, taking into account the nonrigidity of the structure and nonlinearity of the classes.

Algorithms were developed for nonautonomous systems.

The second group of methods deals with decomposition and aggregation relating to the ideas of Bellman and Bailey. For interconnected (large-scale) systems

$$x = (x_1, \ldots, x_m) \in R^n \qquad x_i \in R^{n_i} \left(\sum_{i=1}^{m} n_i = n \right)$$

$$\frac{d\mathbf{x}_i}{dt} = X_i^0(t, x_i) + \sum_{\substack{j=1 \\ j \neq i}}^{n} X_{ij}(t, \mathbf{x})\mathbf{x}_j \qquad (i = \overline{1,m}) \tag{5}$$

Bailey's method of LVF construction is based on:

(1) The decomposition of interconnected system (5) into subsystems

$$\frac{d\mathbf{x}_i}{dt} = X_i^0(t, \mathbf{x}_i) \qquad X_i^0(t, 0) = 0 \qquad \mathbf{x}_i \in R^{n_i} \tag{6}$$

(2) The construction of Ljapunov–Krasovski functions for subsystems (on the assumption of their exponential stability, $\xi = 2$)

$$c_{i1} ||\mathbf{x}_i||^{\xi} \le v^i(t, \mathbf{x}_i) \le c_{i2} ||x_i||^{\xi}$$
$$\dot{v}_{(6)}(t, \mathbf{x}_i) \le c_{i3} ||x_i||^{\xi} (c_{ij} > 0) \tag{7}$$

$$v_i \leq \zeta_i v_i$$

(3) The formation of LVF, $v = (v^1, \ldots, v^m)$ from them
(4) The consideration of the derivative $\dot{v}(t, x)$ using the complete system (5) and taking the interconnections $\mathbf{X}_{ij}(t, \mathbf{x})\mathbf{x}_j$ into account.
(5) A comparison of the construction of the linear system $\dot{\mathbf{x}}_c = P\mathbf{x}_c$, estimating the interconnection between subsystems (aggregation).

This method was applied to multiconnected EES and to nonlinear control systems (NCS). An advantage of this method is its extreme simplicity. A disadvantage lies in "too sufficient" conditions of stability. This concerns a possible roughness of the choice of the P-matrix, the Ljapunov function for subsystems and estimations (7), and a possible failure of system decomposition.

This method was modified by Pioncovski, Zemlyakov, and Furassov *et al.* for a part of the P-matrix choice and $\xi = 1$.

For the elimination of the possible roughness of the P-matrix choice an optimization of the P-matrix choice was used for linear CS, originating from the criteria $\lambda_{\max}(P) \underset{P}{\longrightarrow} \min$ or $|\det P| \underset{P}{\longrightarrow} \max$ under the conditions $p_{ij} \geq 0$ $(j \neq i)$, $Pv - \dot{v}_{(5)} \geq 0$. The algorithm for solving nonlinear programming problems in the linear case was carried out using a computer. Having found P by approximation, we chose $\tilde{f}(v)$ in a quadratic form. CS represents the Riccatty vector equation

$$\dot{x}_c = Px + \operatorname{col} x_c^T Q^s x_c$$

For interconnected systems with linear and polynomial right-hand sides, Abdullin developed a method of LVF and CS construction in which the optimization of P_i-strings of the P-matrix was achieved by a minimization of the differences $\tilde{f}^i(v) - \dot{v}^i(x)$ or of the convex functionals of them.

A decomposition–aggregation method was elaborated and described in many papers by Siljak, and Grujic *et al.* It was applied to EES, the large space telescope, and other ecological and economic systems. This method was strongly developed by Michel *et al.*, Bitsoris, and Furassov *et al.* A method of LVF construction with components as moduli of linear forms was implemented for the parametric control synthesis of interconnected systems by the minimization of the γ_c estimation in the problem of exponential boundedness.

For the multiconnected NCS, Malikov obtained algorithms for the LVF construction with components of Lur'e type, moduli of linear forms, estimations of regions of attraction, and indices of functioning precision.

The algorithms for construction of the region of attraction boundary are based on a combination of the LVF method and an integration of the initial methods. These algorithms were used for an analysis of the absolute stability of multiconnected NCS and for estimations of regions of attraction in EES.

Algorithms were developed for the construction of the LVF with components of quadratic form or as moduli of linear forms, and a comparison of

linear and Rikkatty systems for nonlinear interconnected systems with lag was made:

$$\dot{x}_i(t) = F_i[t, x_i(t), x_i(t-\tau)] + G_i[t, \tilde{x}(t), \tilde{x}(t-\tau)] \qquad (\tau > 0)$$

They were implemented in polynomial right-hand sides and especially in bilinear difference-differential systems. These algorithms were applied in the analysis of exponential stability and regions of attraction of stationary solutions in Marchuk's mathematical models for immunology.

The special iterative process for the decomposition–aggregation method was developed at our institute.

As a first approximation for the LVF and the subsystems we chose the linearized subsystems, the LVF described above, and the linear CS with positive P-matrix, obtained from optimization or by constructing the CS of Rikkatty type. The finite iterative processes for the LVF and CS construction are very effective in applications.

The finite iterative processes for improving the decomposition and construction of LVF and CS were implemented in the study of the stability and dynamics of the first Soviet stratospheric astronomic observatories as well as for other stratospheric and orbital astronomy observatories, i.e., the orbital observatory begun on the space station *Salut-6*. The principal difficulties concerned the high precision of stabilization in space of the observatories on vibrating bases. The results of these investigations were used successfully in practical design (Matrosov *et al.*, 1984).

On the basis of the algorithms that were worked out, a program package for the numerical analysis of exponential stability and other dynamic properties and for the parametric control synthesis of NCS

$$\dot{x} = Ax + By(\delta) + F(t, x) \qquad \sigma = Cx \qquad x \in R^n$$
$$\{x(t+1) = Ax(t) + F_t[x(t)]\}$$

was produced at the Irkutsk Computer Center of the Siberian division of the USSR Academy of Sciences. The programs of this package were tested with success and included in the state fund of algorithms and programs.

This program package was realized in both ALGOL-GDR and FORTRAN-IV as an interactive system, and the man–machine dialogue can be carried out in standard terms of stability, dynamic systems and control theory, and differential and difference equations.

References

Matrosov, V.M. (1962), On the theory of stability, *Appl. Math. Mech.*, **6**.

Matrosov, V.M. (1971), Ljapunov functions in the analysis of nonlinear interconnected systems, *Symp. Math.* V., **6**, pp. 202–242 (Academic Press, New York,

London).

Matrosov, V.M. (1973), Comparison method in systems dynamics, *Proceedings "Equa Diff-73"*, Paris.

Matrosov, V. M., Vassiliev, S.N., *et al.* (1981), *The Algorithms of Theorem Derivations by the Ljapunov Vector Functions Method* (Nauka, Novosibirsk). (In Russian).

Matrosov, V.M. *et al.* (1984), *The Ljapunov Functions Method and Its Applications* (Nauka, Novosibirsk). (In Russian).

Vassiliev, S.N. (1979), The derivation of theorem on dynamic properties, *Algorithm Problem. Algebr. Syst. and Computers*, pp. 3–35 (Irkutsk University, Irkutsk). (In Russian).

Permanence for Replicator Equations

J. Hofbauer[1] and K. Sigmund[2]
[1]*Institute for Mathematics, University of Vienna, Austria*
[2]*International Institute for Applied Systems Analysis, Laxenburg, Austria*

1. Introduction

Many dynamical systems display strange attractors and hence orbits that are so sensitive to initial conditions as to make any long-term prediction (except on a statistical basis) a hopeless task. Such a lack of Ljapunov stability is not always crucial, however: Lagrange stability may be more relevant. Thus, for some models the precise asymptotic behavior – whether it settles down to an equilibrium or keeps oscillating in a regular or irregular fashion - is less important than the fact that all orbits wind up in some preassigned bounded set. The former problem can be impossibly hard to solve and the latter one easy to handle.

Permanence is a stability notion of Lagrangian type which (like the related ones of strong or weak persistence) applies especially well to population dynamical systems, where questions of survival and extinction occur.

The dynamics will be of the form

$$\dot{x}_i = x_i f_i(\mathbf{x}) \tag{1}$$

on $\mathbf{R}^n_+$, or

$$\dot{x}_i = x_i[f_i(\mathbf{x}) - \bar{f}] \tag{2}$$

on the simplex

$$S_n = \{ \mathbf{x} \in \mathbf{R}^n_+ : \sum x_i = 1 \}$$

where

$$\bar{f} = \sum x_i f_i(\mathbf{x}) \tag{3}$$

guarantees that the vector field is tangent to S_n. The variables x_i are

densities or relative frequencies of replicating populations. Such equations describe the effect of selection in a wide variety of fields in theoretical biology, e.g., ecology, genetics, evolutionary game theory, or chemical kinetics (for a survey see Sigmund, 1985).

The boundary of the state space (where some x_i vanish) is invariant. So is the interior, where all types are present. If an orbit in the interior converges to the boundary, this spells extinction for one or several types. The system is called *permanent* if there exists a compact set K in the interior such that all orbits in the interior end up in K. This means that the boundary is a repellor (for $\mathbf{R}^n_+$, we consider the points at infinity as part of the boundary). Equivalently, permanence means that there exists a $k > 0$ such that

$$k < \liminf_{t \to +\infty} x_i(t) \tag{4}$$

for all i, whenever $x_i(0) > 0$ for all i [and for equation (1), in addition, that

$$\limsup_{t \to +\infty} x_i(t) < \frac{1}{k} \tag{5}$$

for all i].

Condition (5) means that orbits are uniformly bounded for $t \to +\infty$, a minimal concession to reality. Condition (4) means that if all types are initially present, selection will not lead to extinction. Even a series of small (but infrequent) perturbations will not be able to wipe out any type. Conversely, if some originally missing component is introduced through mutation, it will spread. The "threshold" k is a uniform one, independent of the initial condition. Thus, permanence is a more stringent property than strong persistence [which requires condition (4) with $k = 0$] and persistence [which requires

$$\limsup x_i(t) > 0$$

for all orbits in the interior of the state space]. Permanence was introduced in Schuster *et al.* (1979). For related stability concepts, we refer to Svirezev and Logofet (1983) and Butler *et al.* (1985).

There are basically only one necessary and one sufficient condition for permanence known so far: we describe them in Section 2. But for most examples, the terms f_i in equations (1) and (2) are linear (see Schuster and Sigmund, 1983, and Hofbauer and Sigmund, 1984). This yields

$$\dot{x}_i = x_i[r_i - (A\mathbf{x})_i] \tag{6}$$

resp.

$$\dot{x}_i = x_i[(A\mathbf{x})_i - \mathbf{x}\cdot A\mathbf{x}] \tag{7}$$

with

$$(A\mathbf{x})_i = \sum a_{ij}x_j \qquad \text{and} \qquad \mathbf{x}\cdot A\mathbf{x} = \sum x_i(A\mathbf{x})_i \,.$$

Equation (6) is the general Lotka–Volterra equation and (7) is the game dynamical equation of Taylor and Jonker (1978) (they are equivalent, as shown in Hofbauer, 1981a). A lot is known on permanence for equations (6) and (7). In Section 3, we present two sufficient conditions, one based on linear inequalities, the other on a geometric feature. Section 4 deals with necessary conditions, all involving the (unique) interior equilibrium. The remainder of this paper discusses two classes of examples: "catalytic networks" in Section 5 and "essentially hypercyclic systems" in Section 6.

It is a pleasure to thank V. Hutson, W. Jansen, E. Amann, and G. Kirlinger for unpublished material and helpful discussions.

2. Fixed Points and Average Ljapunov Functions

Theorem 1 (Hofbauer, 1981b)

Equation (2) is permanent if there exists a function $P : S_n \to \mathbf{R}$ with $P(\mathbf{x}) > 0$ for $\mathbf{x} \in \operatorname{int} S_n$, $P(\mathbf{x}) = 0$ for $\mathbf{x} \in \operatorname{bd} S_n$ and a continuous function $\Psi : S_n \to \mathbf{R}$ such that the following two conditions hold:

(1) For $\mathbf{x} \in \operatorname{int} S_n$,

$$\frac{\dot{P}(x)}{P(x)} = \Psi(\mathbf{x}) \tag{8}$$

(2) For $\mathbf{x} \in \operatorname{bd} S_n$,

$$\frac{1}{T}\int_0^T \Psi[\mathbf{x}(t)]\mathrm{d}t > 0 \qquad \text{for some} \qquad T > 0 \tag{9}$$

The value $P(\mathbf{x})$ measures the distance from $\mathbf{x}$ to the boundary. If one had $\Psi > 0$ on $\operatorname{bd} S_n$ – a condition implying (9) – then $\dot{P}(\mathbf{x}) > 0$ for any $\mathbf{x} \in \operatorname{int} S_n$ near the boundary, and so P would increase, i.e., the orbit would be repelled from $\operatorname{bd} S_n$. In such a case, P would act like a Ljapunov function. Quite often, however, one cannot find a function P of this type. The weaker version defined above is called an average Ljapunov function: it acts in the time average like a Ljapunov function. In the vicinity of the boundary, an orbit need not always move away from the boundary, but it does so in the

mean. For the proof, we refer to Hofbauer (1981b). As an example, it is shown there that

$$P(\mathbf{x}) = x_1 x_2 \cdots x_n \tag{10}$$

is an average Ljapunov function for

$$\dot{x}_i = x_i[x_{i-1}g_i(\mathbf{x}) - \bar{f}] \tag{11}$$

if $g_i(\mathbf{x}) > 0$ for all i and all $\mathbf{x} \in S_n$ (this is the so-called generalized hypercycle, see Hofbauer *et al.*, 1981) and that

$$P(\mathbf{x}) = \prod x_i^{(k_i^{-1})} \tag{12}$$

is an average Ljapunov function for

$$\dot{x}_i = x_i(b_i + k_i x_{i-1} - \bar{f}) \tag{13}$$

for $b_i, k_i > 0$ provided equation (13) admits a rest point in int S_n.

Theorem 1 holds also for equation (1) if its orbits are uniformly bounded.

In Hutson and Moran (1982) and Hutson (1984) this theorem is extended to more general continuous and discrete dynamical systems. It is also shown that it suffices that condition (9) holds for all ω-limit points $\mathbf{x} \in \text{bd}\, S_n$. In Hutson (1986) it is shown that, conversely, every permanent system admits an average Ljapunov function.

Theorem 2 (Hofbauer, 1986)

If equation (1) is permanent then the degree of the vector field with respect to any bounded open set U with $\bar{U} \subseteq \text{int}\, \mathbf{R}^n_+$ containing all interior ω-limits is $(-1)^n$. In particular, there exists a rest point in int $\mathbf{R}^n_+$.

Proof

Let $K \subseteq \text{int}\, \mathbf{R}^n_+$ be a compact set containing all ω-limits in its interior. Let $\tau(\mathbf{x})$ be the time of first entrance into int K;

$$\tau(\mathbf{x}) = \inf\{t \geq 0 : \mathbf{x}(t) \in \text{int}\, K\}$$

It is easy to see that τ is finite on int $\mathbf{R}^n_+$, upper semicontinuous, and therefore locally bounded. Let

$$T \doteq \max_{\mathbf{x} \in K} \tau(\mathbf{x})$$

be the maximum time for orbits leaving K to return to K. The set

$$K^+ = \{\mathbf{x}(t) : \mathbf{x} \in K\,,\ 0 \leq t \leq T\}$$

is compact and forward invariant.

Now let $B \subseteq \operatorname{int} \mathbf{R}^n_+$ be homeomorphic to a ball and contain K^+. The entrance time $\tau(\mathbf{x})$ will again attain an upper bound T^1 on B.

Let $h(\mathbf{x})$ denote the right-hand side of equation (1) and consider the following homotopy:

$$\begin{aligned} h(\mathbf{x},t) &= h(\mathbf{x}) \qquad \text{for} \qquad t = 0 \\ &= \frac{\mathbf{x}(t) - \mathbf{x}(0)}{t} \qquad \text{for} \qquad t > 0 \end{aligned} \tag{14}$$

Clearly, $h(\mathbf{x}\,,\,t) \neq 0$ for all $\mathbf{x} \in \operatorname{bd} B\,,\ t \in \mathbf{R}^+$ since there are no fixed or periodic points on $\operatorname{bd} B$. Thus, the degree of the vector field $\mathbf{x} \to h(\mathbf{x}\,,\,t)$ with respect to B is defined for all $t \in \mathbf{R}$ and independent of t. For $t \geq T^1$, the vector field $h(\mathbf{x}\,,\,t)$ points inward along the boundary of B, and so its degree is $(-1)^n$. Thus, the degree of $h(\mathbf{x})$ with respect to B (or any other open bounded set containing K) is $(-1)^n$.

The same result holds obviously for equation (2). The existence of an interior fixed point for permanent dynamical systems was first proved by Hutson and Vickers (1983) and Sieveking (1983).

3. Properties of the Internal Equilibrium

Theorem 3

If equation (6) is permanent, there exists a unique rest point $\hat{\mathbf{x}}$ in $\operatorname{int} \mathbf{R}^n_+$. For each $\mathbf{x} \in \operatorname{int} S_n$, the time averages

$$\mathbf{z}(T) \doteq \frac{1}{T} \int_0^T \mathbf{x}(t)\, dt \tag{15}$$

converge to $\hat{\mathbf{x}}$ for $T \to +\infty$.

Proof

By *Theorem 2* there exists at least one rest point in $\operatorname{int} \mathbf{R}^n_+$. Such rest points are the solutions (in $\operatorname{int} \mathbf{R}^n_+$) of the linear equation $r_i = (A\mathbf{x})_i$, $i = 1, \ldots, n$. If there were two of them, the line $\mathbf{l}$ joining them would consist of rest points. Since $\mathbf{l}$ intersects $\operatorname{bd} \mathbf{R}^n_+$, it follows that there are rest points arbitrarily

close to the boundary, a contradiction to permanence. Since in int $\mathbf{R}^n_+$

$$(\log x_i)^{\cdot} = r_i - (A\mathbf{x})_i \tag{16}$$

one obtains by integrating from 0 to T and dividing by T

$$\frac{\log x_i(T) - \log x_i(0)}{T} = r_i - [A\mathbf{z}(T)]_i \tag{17}$$

The left-hand side converges to 0 by conditions (4) and (5). Hence each accumulation point of the right-hand side is a rest point in int $\mathbf{R}^n_+$.

The corresponding result holds for equation (7).

Theorem 4

Let equation (6) be permanent and D the Jacobian at the unique interior rest point $\hat{\mathbf{x}}$. Then

$$(-1)^n \det D > 0 \tag{18}$$

$$\operatorname{tr} D < 0 \tag{19}$$

and

$$\det A > 0 \tag{20}$$

Proof

Since $\hat{\mathbf{x}}$ is the unique solution of $r_i = (A\mathbf{x})_i$, the matrix A is nonsingular. But

$$d_{ij} = -\hat{x}_i a_{ij} \tag{21}$$

and so D is also nonsingular. *Theorem 2* shows that the index $i(\hat{\mathbf{x}})$ is $(-1)^n$. But $i(\hat{\mathbf{x}})$ is just the sign of $\det D$. This and equation (21) imply (18) and (20). In order to prove (19), we multiply the right-hand side of equation (6) with the positive function

$$B(\mathbf{x}) = \Pi\, x_i^{s_i - 1} \tag{22}$$

The resulting equation $\dot{x}_i = h_i(\mathbf{x})$, with

$$h_i(\mathbf{x}) \doteq x_i \left[r_i - \sum a_{ij} x_j \right] B(\mathbf{x}) \tag{23}$$

differs from equation (6) just by a change in velocity and is therefore also permanent. Since

$$\frac{\partial h_i}{\partial x_i} = B(\mathbf{x})\{ s_i[r_i - (A\mathbf{x})_i] - x_i a_{ii} \} \tag{24}$$

we obtain

$$\operatorname{div} \mathbf{h}(\mathbf{x}) = B(\mathbf{x})\left\{\sum s_i[r_i - (A\mathbf{x})_i] - \sum x_i a_{ii}\right\} \tag{25}$$

which at the rest point $\hat{\mathbf{x}}$ reduces to

$$\operatorname{div} \mathbf{h}(\hat{\mathbf{x}}) = -B(\hat{\mathbf{x}}) \sum \hat{x}_i\, a_{ii} = B(\hat{\mathbf{x}}) \operatorname{tr} D \tag{26}$$

Hence

$$\begin{aligned}\operatorname{div} \mathbf{h}(\mathbf{x}) &= B(\mathbf{x}) \left\{ \sum_i s_i \left[\sum_j a_{ij}(\hat{x}_j - x_j)\right] + \sum a_{ii}(\hat{x}_i - x_i) + \operatorname{tr} D\right\} \\ &= B(\mathbf{x}) \left[\sum_j (\hat{x}_j - x_j)\left(\sum_i s_i\, a_{ij} + a_{jj}\right) + \operatorname{tr} D\right]\end{aligned}$$

Since A is nonsingular, we may choose s_i such that

$$\sum_i s_i\, a_{ij} + a_{jj} = 0 \tag{27}$$

Then

$$\operatorname{div} \mathbf{h}(\mathbf{x}) = B(\mathbf{x}) \operatorname{tr} D \tag{28}$$

Since there exists a ball in $\operatorname{int} \mathbf{R}^n_+$ which, in time T, shrinks (see the proof of *Theorem 2*), Liouville's theorem and equation (28) imply (20).

Similarly, if equation (7) is permanent and D denotes the Jacobian at the unique interior rest point $\hat{\mathbf{x}}$ then

$$(-1)^{n-1} \det D > 0 \tag{29}$$

$$\operatorname{tr} D < 0 \tag{30}$$

and if $a_{ii} = 0$ for all i

$$(-1)^{n-1} \det A > 0 \tag{31}$$

[We note that equation (7) is unchanged in S_n if one adds constants to each column of A; hence it can always be realized by matrices A with zero diagonal.] Indeed, conditions (29) and (30) follow from (18) and (19) through the equivalence of equations (6) and (7). In particular, Hofbauer's proof (1981a) shows that equation (7) can be transformed by a smooth change in coordinates, followed by a change in velocity, into the equation

$$\dot{x}_i = x_i[r_i - (A'\mathbf{x})_i] \tag{32}$$

in $n-1$ variables, with $a'_{ij} = a_{nj} - a_{ij}$. Now a simple computation yields

$$\operatorname{tr} D = \sum_i a_{ii}\hat{x}_i - \hat{\mathbf{x}} \cdot A\hat{\mathbf{x}} \tag{33}$$

Thus, by condition (30)

$$\hat{\mathbf{x}} \cdot A\hat{\mathbf{x}} > 0 \tag{34}$$

Furthermore, since

$$(A\hat{\mathbf{x}})_i = \hat{\mathbf{x}}A\hat{\mathbf{x}}$$

for all i, Cramer's rule implies

$$\hat{x}_n \det A = (\hat{\mathbf{x}} \cdot A\hat{\mathbf{x}}) \det A_n$$

with

$$\det A_n = \begin{bmatrix} a_{11} & \cdots & a_{1,n-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 1 \end{bmatrix}$$

Clearly

$$\det A_n \equiv \begin{bmatrix} -a'_{11} & \cdots & -a'_{1,n-1} & 0 \\ \vdots & & \vdots & \vdots \\ -a'_{n-1,1} & \cdots & -a'_{n-1,n-1} & 0 \\ a_{n,1} & \cdots & a_{n,n-1} & 1 \end{bmatrix} = (-1)^{n-1} \det A'$$

with A' as in equation (32). Since this is permanent, $\det A' > 0$, which together with condition (34) implies (31).

Before proceeding further, it will be useful to define a rest point $\mathbf{p}$ of equation (6) [resp. (7)] as *saturated* if $r_i \leq (A\mathbf{p})_i$ [resp. $(A\mathbf{p}]_i \leq \mathbf{p} \cdot A\mathbf{p}$] whenever $p_i = 0$ (note that the equality sign must hold whenever $p_i > 0$). Every rest point in the interior of the state space is trivially saturated. For a rest point on the boundary, the condition means that selection does not "call for" the missing species.

Theorem 5 (Hofbauer, 1986)

Equation (7) has at least one saturated rest point. If all the saturated rest points have nonsingular Jacobian, the sum of their indices is $(-1)^{n-1}$ (and hence their number is odd).

Proof

Let us suppose first that all saturated rest points $\mathbf{p}$ have nonsingular Jacobian. Then $(A\mathbf{p})_i < \mathbf{p} \cdot A\mathbf{p}$ whenever $p_i = 0$, since the $(A\mathbf{p})_i - \mathbf{p} \cdot A\mathbf{p}$ are eigenvalues of $\mathbf{p}$. Let us consider

$$\dot{x}_i = x_i\,[(A\mathbf{x})_i - \mathbf{x} \cdot A\mathbf{x} - n\varepsilon] + \varepsilon \tag{35}$$

as a perturbation of equation (7). (The small $\varepsilon > 0$ represents biologically an immigration.) Clearly, equation (35) maintains $\sum \dot{x}_i = 0$ on S_n. On bd S_n, the flow now points into int S_n. The sum of the indices of the rest points of equation (35) in int S_n is therefore $(-1)^{n-1}$. They correspond to the saturated rest points of equation (7). Indeed the rest points $\mathbf{p} = \mathbf{p}(\varepsilon)$ of equation (35) satisfy

$$p_i(\varepsilon) = \frac{\varepsilon}{n\varepsilon + \mathbf{p} \cdot A\mathbf{p} - (A\mathbf{p})_i}$$

If $\mathbf{p}(0)$ is saturated, then $(A\mathbf{p})_i < \mathbf{p} \cdot A\mathbf{p}$ and hence $p_i(\varepsilon) > 0$. If $\mathbf{p}(0)$ is not saturated, $(A\mathbf{p})_i > \mathbf{p} \cdot A\mathbf{p}$ for some i and so $p_i(\varepsilon) < 0$. Now small perturbations leave the Jacobian nonsingular and do not change the index (which is +1 or −1). This proves the second part.

Now let us drop the regularity assumption. Since equation (35) has index sum $(-1)^{n-1}$, it admits at least one rest point $\mathbf{p}(\varepsilon)$ in int S_n. Let $\mathbf{p}$ be an accumulation point of $\mathbf{p}(\varepsilon)$, for $\varepsilon \to 0^+$. The previous equation implies

$$n\varepsilon + \mathbf{p}(\varepsilon) \cdot A\mathbf{p}(\varepsilon) - [A\mathbf{p}(\varepsilon)]_i > 0$$

for all i, and thus $\mathbf{p} \cdot A\mathbf{p} \geq (A\mathbf{p})_i$ by continuity. Thus $\mathbf{p}$ is a saturated rest point in S_n.

A similar result holds for equation (6) under the assumption of uniform boundedness of all orbits.

The saturated rest points $\mathbf{p}$ of equation (7) are the Nash equilibria for the symmetric game with payoff matrix A:

$$\mathbf{x} \cdot A\mathbf{p} \leq \mathbf{p} \cdot A\mathbf{p} \tag{36}$$

for all $\mathbf{x} \in S_n$ (the strategy $\mathbf{p}$ is a best reply to itself). Thus one obtains, as immediate corollaries of *Theorem 5*, the existence of a Nash equilibrium and the odd number theorem for regular Nash equilibria, both classical results in game theory (we have considered only symmetric games, but the same dynamical proof works for bimatrix games too).

We shall call equation (7) robustly persistent if it remains persistent under small perturbations of the a_{ij}.

Theorem 6

If equation (7) is robustly persistent, then it has a unique interior rest point $\hat{\mathbf{x}}$ and its index is $(-1)^{n-1}$.

Proof

In a robustly persistent system there are no saturated rest points on the boundary. Indeed, if $\bar{\mathbf{x}}$ were such a rest point with $(A\bar{\mathbf{x}})_i \leq \bar{\mathbf{x}} \cdot A\bar{\mathbf{x}}$ for all i with $\bar{x}_i = 0$, then a suitable perturbation would turn $\bar{\mathbf{x}}$ into a hyperbolic fixed point with all external eigenvalues negative. But then the stable manifold of $\bar{\mathbf{x}}$ meets int S_n. Thus, there are interior orbits converging to $\bar{\mathbf{x}}$, contradicting persistence.

Since there exists at least one saturated fixed point, equation (7) has an interior equilibrium and its index is $(-1)^{n-1}$.

4. Sufficient Conditions for Permanence

Theorem 7 (Jansen, 1986)

Equation (7) is permanent if there exists a $\mathbf{p} \in \text{int } S_n$ such that

$$\mathbf{p} \cdot A\mathbf{x} > \mathbf{x} \cdot A\mathbf{x} \tag{37}$$

holds for all rest points $\mathbf{x}$ on bd S_n.

Proof

We show that

$$P(\mathbf{x}) = \Pi\, x_i^{p_i} \tag{38}$$

is an average Ljapunov function. Clearly $P(\mathbf{x}) = 0$ for $\mathbf{x} \in \text{bd}\, S_n$ and $P(\mathbf{x}) > 0$ for $\mathbf{x} \in \text{int}\, S_n$. The function

$$\Psi(\mathbf{x}) = \mathbf{p} \cdot A\mathbf{x} - \mathbf{x} \cdot A\mathbf{x} \tag{39}$$

satisfies $\dot{P} = P\Psi$ in int S_n. It remains to show that for every $\mathbf{y} \in \text{bd}\, S_n$, there is a $T > 0$ such that

$$\int_0^T \Psi[\mathbf{y}(t)]\, dt > 0 \tag{40}$$

The proof proceeds by induction on the number r of positive components of $\mathbf{y}$. For $r = 1$, $\mathbf{y}$ is a corner of S_n and hence a rest point, so that condition (40) is an obvious consequence of condition (37). Assume now that (40) is valid for $r \leq m - 1$ and that

$$I = \{i : 1 \leq i \leq n\,,\ y_i > 0\}$$

has cardinality m. We have to distinguish two cases.

Case I. $\mathbf{y}(t)$ converges to the boundary of the simplex

$$S(I) = \{\mathbf{x} \in S_n : x_i = 0 \text{ for all } i \notin I\}$$

Since bd $S(I)$ consists of faces of dimension $\leq m - 1$, our assumption implies that for every $\bar{y} \in \text{bd}\, S(I)$, there exists a $T(\bar{\mathbf{y}}) \geq 1$ and an $\alpha(\bar{y}) > 0$ such that

$$\int_0^{T(\bar{\mathbf{y}})} \Psi[\bar{\mathbf{y}}(t)]\, dt > \alpha(\bar{\mathbf{y}})$$

and even that

$$\int_0^{T(\bar{\mathbf{y}})} \Psi[\mathbf{x}(t)]\, dt > \alpha(\bar{\mathbf{y}})$$

for all $\mathbf{x}$ in some neighborhood $U(\bar{\mathbf{y}})$. Since bd $S(I)$ is compact, we may cover it by finitely many $U(\bar{\mathbf{y}}_k)$: their union U contains $\mathbf{y}(t)$ for all t larger than some T_0. Now $\mathbf{y}(T_0) \in U$ implies

$$\int_{T_0}^{T_1} \Psi[\mathbf{y}(t)]\,dt > \alpha$$

for some $T_1 > T_0$ [where α is the minimum of the $\alpha(\bar{\mathbf{y}}_k)$]. Similarly $\mathbf{y}(T_1) \in U$ implies

$$\int_{T_1}^{T_2} \Psi[\mathbf{y}(t)]\,dt > \alpha$$

etc. Now

$$\int_0^{T_s} \Psi[\mathbf{y}(t)]\,dt > \int_0^{T_0} \Psi[\mathbf{y}(t)]\,dt + \alpha s$$

and this number is positive if s is large enough.

Case II. $\mathbf{y}(t)$ does not converge to bd $S(I)$.

In this case there exists an $\varepsilon > 0$ and a sequence $T_k \to +\infty$ such that this $y_i(T_k) > \varepsilon$ for *all* $i \in I$ and $k = 1,2,\ldots$.

Let us write

$$x_i(T) \doteq \frac{1}{T}\int_0^T y_i(t)\,dt$$

and

$$a(T) \doteq \frac{1}{T}\int_0^T \mathbf{y}(t) \cdot A\,\mathbf{y}(t)\,dt$$

Since the sequences $x_i(T_k)$ and $a(T_k)$ are bounded, we may obtain a subsequence – which we again denote by T_k – such that $x_i(T_k)$ and $a(T_k)$ converge for all i: the limits will be denoted by $\bar{x}_i$ and $\bar{a}$. For $i \in I$, we have $y_i(t) > 0$ and hence

$$(\log y_i)^{\cdot} = (A\,\mathbf{y})_i - \mathbf{y} \cdot A\,\mathbf{y}$$

Integrating this from 0 to T_k and dividing by T_k, we obtain

$$\frac{1}{T_k}[\log y_i(T_k) - \log y_i(0)] = [A\mathbf{x}(T_k)]_i - a(T_k)$$

Since $\log y_i(T_k)$ is bounded, the left-hand side converges to 0. Hence

$$(A\bar{\mathbf{x}})_i = \bar{a} \quad \text{for } i \in I$$

Furthermore $\sum \bar{x}_i = 1$, $\bar{x}_i \geq 0$, and $\bar{x}_i = 0$ for $i \notin I$. Therefore $\bar{\mathbf{x}}$ is an equilibrium in $S(I)$ and $\bar{a} = \bar{\mathbf{x}} \cdot A\bar{\mathbf{x}}$. Now

$$\frac{1}{T_k}\int_0^{T_k} \Psi[\mathbf{y}(t)]\,dt = \sum_{i=1}^{n} p_i \frac{1}{T_k}\int_0^{T_k} [(A\mathbf{y})_i - \mathbf{y} \cdot A\mathbf{y}]\,dt$$

converges to

$$\sum_{i=1}^{n} p_i\,[(A\bar{\mathbf{x}})_i - \bar{\mathbf{x}} \cdot A\bar{\mathbf{x}}] = \mathbf{p} \cdot A\bar{\mathbf{x}} - \bar{\mathbf{x}} \cdot A\bar{\mathbf{x}}$$

which is strictly positive by equation (37).

Thus, in order to prove permanence, one has to find out whether there exists a positive solution $\mathbf{p}$ for the linear inequalities

$$\sum_{i:x_i=0} p_i[(A\mathbf{x})_i - \mathbf{x} \cdot A\mathbf{x}] > 0 \tag{41}$$

(where $\mathbf{x}$ runs through the boundary rest points). The coefficients $(A\mathbf{x})_i - \mathbf{x} \cdot A\mathbf{x}$ are the eigenvalues of $\mathbf{x}$ whose eigenvectors are transversal to the boundary face of $\mathbf{x}$. The number of inequalities may be quite large: but it is easy to check that if continua of rest points occur, the inequalities (41) have to be tested for extreme points only, so that the problem is finite.

It is remarkable that only the rest points are involved: the ω-limit sets on the boundary may be considerably more complicated.

Similarly, an equation of type (6) with uniformly bounded orbits is permanent if there exists a positive solution $\mathbf{p}$ for

$$\sum_{i:x_i=0} p_i[r_i - (A\mathbf{x})_i] > 0 \tag{42}$$

(where $\mathbf{x}$ runs through the boundary rest points).

Theorem 8 (Hofbauer, 1986)

Consider equation (6) with uniformly bounded orbits. If the set

$$D = \{\mathbf{x} \in \mathbf{R}^n_+ : \mathbf{r} \leq A\mathbf{x}\} \tag{43}$$

(where no species increases) is disjoint from the convex hull C of all boundary fixed points, then equation (6) is permanent.

Proof

A rest point $\mathbf{z}$ of equation (6) lies in D if and only if it is saturated. The assumption $C \cap D = \phi$ implies that there are no saturated rest points on the boundary. *Theorem 5* shows then the existence of an interior rest point $\hat{\mathbf{x}}$; this point is unique, since otherwise we would have a line of fixed points intersecting bd $\mathbf{R}^n_+$, i.e., a nonempty intersection of C and D. It follows that A is nonsingular.

The convex set C can be separated from the convex set $\hat{D} = \{\mathbf{x} \in \mathbf{R}^n : \mathbf{r} \leq A\mathbf{x}\}$ by a hyperplane. Since A is nonsingular we can write the separating functional in the form $\mathbf{p}A$, with $\mathbf{p} \in \mathbf{R}^n$. Then

$$\mathbf{p} \cdot A\mathbf{z} < \mathbf{p} \cdot A\mathbf{x} \tag{44}$$

for all $\mathbf{x} \in \hat{D}$ and all fixed points $\mathbf{z} \in$ bd $\mathbf{R}^n_+$. Since the interior fixed point $\hat{\mathbf{x}}$ lies in $\hat{D}$ we obtain in particular

$$\mathbf{p} \cdot A\mathbf{z} < \mathbf{p} \cdot A\hat{\mathbf{x}} = \mathbf{p} \cdot \mathbf{r} \tag{45}$$

Thus $\mathbf{p}$ is a solution of equation (42). We are left to show that $\mathbf{p}$ is positive.

Let $\mathbf{v}$ be any vector in $\mathbf{R}^n$ with $A\mathbf{v} > \mathbf{0}$. Then $\hat{\mathbf{x}} + c\ \mathbf{v}$ belongs to $\hat{D}$ for every $c > 0$ since

$$A(\hat{\mathbf{x}} + c\ \mathbf{v}) = \mathbf{r} + c\ A\mathbf{v} > \mathbf{r}$$

Thus

$$\mathbf{p} \cdot A\ \mathbf{z} < \mathbf{p} \cdot A\hat{\mathbf{x}} + c\ \mathbf{p} \cdot A\mathbf{v} \tag{46}$$

This can hold for arbitrarily large c only if $\mathbf{p} \cdot A\mathbf{v} \geq 0$ for all such $\mathbf{v}$. Therefore $\mathbf{p} \geq \mathbf{0}$. Since the set of solutions $\mathbf{p}$ of (42) is open and $\mathbf{z}$ varies in a compact set, we can find a solution $\mathbf{p} > \mathbf{0}$.

5. Catalytic Networks

To equation (7) with $a_{ij} \geq 0$ for all i and j, we associate an oriented graph: an arrow from j to i denotes that $a_{ij} > 0$, i.e., that j enhances the replication of i.

A directed graph is called irreducible if for any two species i and j there exists an oriented path leading from i to j.

Theorem 9

If equation (7) with $a_{ij} \geq 0$ is permanent, then its graph is irreducible.

We refer to Schuster and Sigmund (1984) for a proof.

By a catalytic network, we understand an equation of type (7) with $a_{ij} \geq 0$ and $a_{ii} = 0$ for all i.

Theorem 10 (Amann, 1986)

For $n \leq 5$, the graph of a permanent catalytic network is Hamiltonian, i.e., contains a closed path visiting each vertex exactly once.

Proof

We know from condition (31) that the sign of det A is $(-1)^{n-1}$. Now

$$\det A = \sum_{\sigma} \operatorname{sgn} \sigma \; a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where the summation extends over all permutations σ of $\{1, \cdots, n\}$. Since $a_{ii} = 0$ we need only consider permutations without fixed elements. Every permutation σ can be split into some k elementary cycles, and $\operatorname{sgn} \sigma = (-1)^{n-k}$. If $n \leq 5$, k can be 1 or 2. Any permutation that is the product of two smaller cycles has sign $(-1)^{n-2}$. In order that $\operatorname{sgn} \det A = (-1)^{n-1}$, there must be at least one permutation σ consisting of a single cycle such that

$$a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} > 0 \tag{47}$$

This corresponds to a closed feedback loop visiting every vertex precisely once.

Amann (1986) shows that for $n \geq 6$ a catalytic network may be permanent even if its graph is not Hamiltonian. For $n = 6$, his example is

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 3 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{48}$$

Indeed, condition (37) is satisfied with $\mathbf{p} = \frac{1}{49}(9,1,2,13,23,1)$. The inequalities corresponding to boundary fixed points $\mathbf{x}$ with $\mathbf{x} \cdot A\mathbf{x} = 0$ are trivially

satisfied since A is irreducible. The only remaining fixed points are

$$\left(\frac{1}{3},\frac{2}{3},0,0,0,0\right) \qquad \left(,0,\frac{1}{3},\frac{2}{3},0,0\right) \qquad \left(0,0,0,0,\frac{1}{2},\frac{1}{2}\right)$$

$$\frac{1}{41}(9,6,0,5,18,3) \qquad \frac{1}{28}(5,0,9,2,9,3)$$

It is not difficult to show that for $n = 3$ a catalytic network is permanent if and only if it admits a unique rest point in int S_3. Up one dimension, one has *Theorem 11*.

Theorem 11 (Amann, 1986)

For a catalytic network with $n = 4$, a necessary and sufficient condition for permanence is that there exists an interior fixed point and $\det A < 0$.

Proof

Necessity follows from *Theorem 2* and condition (31). In order to prove sufficiency, we use Jansen's method and look for a positive solution $\mathbf{p}$ of (41). As a first step we show that only those boundary equilibria $\mathbf{x}$ have to be taken into account that lie on an edge and are isolated within that edge:

(1) For the corners of S_n, condition (41) is satisfied for every choice $\mathbf{p} > 0$, since they must have some positive eigenvalue: each column of A contains a positive element.

(2) Next we show that condition (41) is satisfied for every choice $\mathbf{p} > 0$ if $\mathbf{x}$ is a rest point in the interior of the 3-face $x_k = 0$. We need only to consider the case where $\mathbf{x}$ is isolated [indeed, condition (41) need only to be checked at the extreme points of continua of boundary equilibria]. Then $A^{(k)}$, the 3×3-minor of A obtained by deleting its kth row and column, is nonsingular. A straightforward computation shows that the external eigenvalue of $\mathbf{x}$, i.e., $(A\mathbf{x})_k - \mathbf{x}\cdot A\mathbf{x}$, has the same sign as

$$\frac{-(\hat{\mathbf{x}}\cdot A\hat{\mathbf{x}})\det A}{(\mathbf{x}\cdot A\mathbf{x})\det A^{(k)}}$$

Since the restriction of the catalytic network to the face $x_k = 0$ is permanent, we have $\det A^{(k)} > 0$. Hence $(A\mathbf{x})_k > \mathbf{x}\cdot A\mathbf{x}$ and condition (41) is trivially satisfied.

(3) Thus, we have only to check condition (41) for the rest points on edges. Let F_{ij} denote such a point on the edge from $\mathbf{e}_i$ to $\mathbf{e}_j$, and let Γ_{ij}^k be its eigenvalue in the direction $\mathbf{e}_k$. Then

$$\Gamma_{ij}^k = \frac{a_{ki}\,a_{ij} + a_{kj}\,a_{ji} - a_{ij}\,a_{ji}}{a_{ij} + a_{ji}} \tag{49}$$

We note first that F_{ij} cannot be saturated. Indeed, suppose that $\Gamma_{ij}^k \leq 0$ and $\Gamma_{ij}^l \leq 0$ for the two indices k, l different from i and j. Both the conditions on the interior rest point $\hat{\mathbf{x}}$ and the sign of $\det A$ are open. Thus we may, without affecting them, slightly perturb the coefficients such that every fixed point is regular and Γ_{ij}^k, $\Gamma_{ij}^l < 0$. But the index of every saturated regular rest point on an edge is -1. Since $\det A < 0$ implies that the index of $\hat{\mathbf{x}}$ is also -1, this yields a contradiction to *Theorem 4*.

From condition (41) we are left with (up to 6) inequalities of the form

$$p_k \Gamma_{ij}^k + p_l \Gamma_{ij}^l > 0 \tag{50}$$

Now

$$\Gamma_{ij}^k < 0 \Longrightarrow \Gamma_{ik}^k > 0 \qquad \text{and} \qquad \Gamma_{jk}^i > 0 \tag{51}$$

Indeed, $\Gamma_{ij}^k < 0$ implies $a_{ki} < a_{ji}$ and $a_{kj} < a_{ij}$, and hence $\Gamma_{ik}^j > 0$ and $\Gamma_{jk}^i > 0$. Similarly

$$\Gamma_{lk}^i < 0\,, \qquad \Gamma_{jl}^k < 0 \Longrightarrow \Gamma_{il}^j > 0 \tag{52}$$

Indeed, $\Gamma_{lk}^i < 0$ implies $a_{ii} < a_{ki}$ and $\Gamma_{jl}^k < 0$ implies $a_{ki} < a_{ji}$, so that $\Gamma_{il}^j > 0$.

By condition (51) at least two of the inequalities (50) are trivially satisfied. We write $l \to k$ if $\Gamma_{ij}^l < 0$ and $\Gamma_{ij}^k > 0$. If this relation has no cycles, then inequality (50) is easy to solve: whenever $l \to k$, one simply chooses a very large p_k. By (52), there can be no cycles of length 3. So we are left with the case of a cycle of length 4. After suitably rearranging the indices, this means

$$\Gamma_{23}^1\,,\ \Gamma_{34}^2\,,\ \Gamma_{41}^3\,,\Gamma_{12}^4 < 0 \tag{53}$$

while all other Γs are positive by (51). Then the system (50) admits a solution $\mathbf{p} > 0$ if and only if

$$\Gamma_{23}^1 \Gamma_{34}^2 \Gamma_{41}^3 \Gamma_{12}^4 < \Gamma_{34}^1 \Gamma_{41}^2 \Gamma_{12}^3 \Gamma_{23}^4 \tag{54}$$

It remains to show that (54) holds. Now

$$0 < -\Gamma_{23}^1 \frac{a_{23} + a_{32}}{a_{32} - a_{12}} \qquad (a_{32} > a_{12} \text{ since } \Gamma_{23}^1 < 0)$$

$$= a_{23} - \frac{a_{13}\, a_{32}}{a_{32} - a_{12}}$$

$$< a_{23} < a_{43} \qquad \text{(because } \Gamma^2_{34} < 0\text{)}$$

$$< a_{43} + \frac{a_{13}\, a_{32}}{a_{14} - a_{34}} \qquad (a_{14} > a_{34} \text{ since } \Gamma^3_{41} < 0)$$

$$= \Gamma^1_{34} \frac{a_{34} + a_{43}}{a_{14} - a_{34}}$$

Multiplying

$$0 < -\Gamma^1_{23} \frac{a_{23} + a_{32}}{a_{32} - a_{12}} < \Gamma^1_{34} \frac{a_{34} + a_{43}}{a_{14} - a_{34}} \tag{55}$$

with its circular permutations yields (54) and concludes the proof.

6. Essentially Hypercyclic Systems

An $n \times n$-matrix A is called essentially hypercyclic if $a_{ij} > 0$ for $i = j + 1 \pmod n$, $a_{ii} = 0$ and $a_{ij} \leq 0$ for all other i, j. This corresponds to a closed feedback loop of positive interactions, together (possibly) with some negative interactions.

Permanence of such networks can be characterized in terms of M-matrices. An $n \times n$-matrix C with nonpositive diagonal terms $c_{ij} \leq 0$ $(i \neq j)$ is called an M-matrix if one of the following equivalent conditions is satisfied:

(1) $\exists\, \mathbf{p} > 0$ such that $C\mathbf{p} > 0$.
(2) All principal minors of C are positive.
(3) All real eigenvalues of C are positive.
(4) For all $\mathbf{x} > 0$ there is an i such that $(C\mathbf{x})_i > 0$.

Theorem 12

For equation (7) with an essentially hypercyclic matrix A, the following conditions are equivalent:

(1) The system is permanent.
(2) There is an interior equilibrium $\hat{\mathbf{x}}$ where the Jacobian D has negative trace.
(3) There is an interior equilibrium $\hat{\mathbf{x}}$ with $\hat{\mathbf{x}} \cdot A\hat{\mathbf{x}} > 0$.
(4) The matrix C obtained by moving the top row of A to the bottom is an M-matrix.

(5) There is a $\mathbf{p} > 0$ such that $\mathbf{p}A > 0$.

Proof

(1) → (2) has been shown in *Theorem 2* and condition (30).
(2) → (3) is just equation (33).
(3) → (4) since $(A\hat{\mathbf{x}})_i = \hat{\mathbf{x}} \cdot A\hat{\mathbf{x}} > 0$, thus $A\hat{\mathbf{x}} > 0$, and hence $C\hat{\mathbf{x}} > 0$.
(4) → (5) because for some vector $\mathbf{p} > 0$, one has $(p_2, \ldots, p_n, p_1)\, C > 0$.
(5) → (1) follows from the fact that equation (38) is an average Ljapunov function. We only have to note that $\Psi(\mathbf{x}) = \mathbf{p} \cdot A\mathbf{x} - \mathbf{x} \cdot A\mathbf{x}$ satisfies condition (9) in *Theorem 1*. We proceed indirectly. Assume there exists an $\mathbf{x} \in \operatorname{bd} S_n$ such that for all $T > 0$

$$\frac{1}{T}\int_0^T \Psi[\mathbf{x}(t)]\, dt \leq 0$$

i.e.

$$\frac{1}{T}\int_0^T \mathbf{x}(t)\, dt \geq \frac{1}{T}\int_0^T \mathbf{p} \cdot A\mathbf{x}(t)\, dt > \varepsilon \tag{56}$$

for some suitable $\varepsilon > 0$, since $\mathbf{p}A > 0$. This implies $x_i(t) \to 0$ for all i. Indeed, $x_i(t) \to 0$ is satisfied for at least one i (since $\mathbf{x} \in \operatorname{bd} S_n$). Now if $x_{i+1} > 0$, then

$$\frac{\dot{x}_{i+1}}{x_{i+1}} = a_{i+1,i}\, x_i + \sum_{j \neq i, i+1} a_{i+1,j}\, x_j - \mathbf{x} \cdot A\mathbf{x}$$

Integrating from 0 to T and using the fact that the sum on the right-hand side is nonpositive, one obtains

$$\frac{1}{T}\left[\log x_{i+1}(T) - \log x_{i+1}(0)\right] \leq a_{i+1,i}\, \frac{1}{T}\int_0^T x_i(t)\, dt - \frac{1}{T}\int_0^T \mathbf{x}(t) \cdot A\mathbf{x}(t)\, dt$$

The first integral on the right-hand side converges to 0 while the second one is larger than ε by condition (56). Hence $x_{i+1}(T) \leq \exp(-\varepsilon T)$ and so $x_{i+1}(T) \to 0$ for $T \to +\infty$. All components of $\mathbf{x}(t)$ converge to 0, a contradiction to $\mathbf{x}(t) \in S_n$.

Each of the conditions (2), (3), (4), and (5) of *Theorem 12* is easily checked for the hypercycle

$$\dot{x}_i = x_i(k_i x_{i-1} - \bar{f}) \tag{57}$$

with $k_i > 0$. This theorem also yields permanence criteria for

$$\dot{x}_i = x_i(a_{i-1} x_{i-1} + b_{i+1} x_{i+1} - \bar{f}) \tag{58}$$

and

$$\dot{x}_i = x_i(a_i x_i + b_{i-1} x_{i-1} - \bar{f}) \tag{59}$$

(see Amann and Hofbauer, 1985).

7. Discussion

Interesting permanence criteria have been found for low dimensional ecological models. Thus, Hutson and Vickers (1983) have completely characterized two prey–one predator systems described by

$$\begin{aligned}
\dot{x}_1 &= x_1(r_1 - a_{11} x_1 - a_{12} x_2 + b_1 y) \\
\dot{x}_2 &= x_2(r_2 - a_{21} x_2 - a_{22} x_1 + b_2 y) \\
\dot{y} &= y(-r_3 + \beta_1 x_1 + \beta_2 x_2 - \gamma y)
\end{aligned}$$

Such systems are permanent if and only if:

(1) There exists an interior rest point.
(2) $\det A > 0$.
(3) The $(x_1 - x_2)$-subsystem is not bistable [which means that there exists an unstable equilibrium in the interior of the positive $(x_1 - x_2)$-orthant].

In particular, a system of two competing species can be "stabilized" (i.e., can become a permanent 3-system) by the introduction of a suitable predator, if and only if it is not bistable.

Recently, Kirlinger (1986) has characterized permanence for two predator–two prey systems of the form

$$\dot{x}_1 = x_1(r_1 - a_{11}x_1 - a_{12}x_2 - b_1 y_1)$$
$$\dot{x}_2 = x_2(r_2 - a_{21}x_1 - a_{22}x_2 - b_2 y_2)$$
$$\dot{y}_1 = y_1(-r_3 + \beta_1 x_1)$$
$$\dot{y}_2 = y_2(-r_4 + \beta_2 x_2)$$

and shown that bistable $(x_1 - x_2)$-systems can be "stabilized" by the introduction of *two* predators – Kirlinger has also treated two predator–two prey systems with interspecific competition between the predators and three predator–three prey systems. Such couplings of predator–prey systems were first considered by Svirezev and Logofet (1983).

Kirlinger (1986) has also shown that one prey–two predator systems are permanent if and only if there exists an interior rest point and $\det A > 0$.

Persistence of three competing species was characterized by Hallam *et al.* (1979). Permanence and persistence of general three species Lotka–Volterra systems were finally characterized by Butler and Waltman (1984) and Hofbauer (1986).

For permanence results on ecological systems with nonlinear interaction terms, we refer to Hutson (1984), Hutson and Law (1985), and Freedman and Waltman (1985). We also mention Hofbauer (1984), and Hutson and Moran (1982) for the case of difference equations.

References

Amann, E. (1986), *Permanence of Catalytic Networks* (to appear).

Amann, E. and Hofbauer, J. (1985), Permanence in Lotka–Volterra and replicator equations, in E. Ebeling and M. Peschel (Eds), *Lotka–Volterra Approaches to Cooperation and Competition in Dynamic Systems* (Akademie Verlag, Berlin).

Butler, G. and Waltman, P.E. (1984), *Persistence in Three-Dimensional Lotka–Volterra Systems* (to appear).

Butler, G., Freedman, H.I., and Waltman, P.E. (1985), Uniformity persistent systems, to appear in *Proceedings AMS*.

Freedman, H.I. and Waltman, P.E. (1985), Persistence in a model of three competitive populations, *Math. Biosciences*, **73**, 89–101.

Hallam, T., Svoboda, L., and Gard, T. (1979), Persistence and extinction in three species Lotka–Volterra competitive systems, *Math. Biosciences*, **46**, 117–124.

Hofbauer, J. (1981a), On the occurrence of limit cycles in the Volterra–Lotka equation, *Nonlinear Analysis, TMA*, **5**, 1003–1007.

Hofbauer, J. (1981b), A general cooperation theorem for hypercycles, *Monatsh. Math.*, **91**, 233–240.

Hofbauer, J. (1984), A difference equation model for the hypercycle, *SIAM J. Appl. Math.*, **44**, 762–772.

Hofbauer, J. (1986), *Permanence and Persistence of Lotka-Volterra Systems* (to appear).

Hofbauer, J. and Sigmund, K. (1984), *Evolutionstheorie und dynamische Systeme* (Paul Parey Verlag, Hamburg, Berlin).

Hofbauer, J., Schuster, P., and Sigmund, K. (1981), Competition and cooperation in catalytic self-replication, *J. Math. Biol.*, **11**, 155–168.

Hutson, V. (1984), Predator mediated coexistence with a switching predator, *Math. Biosci.*, **63**, 293–269.

Hutson, V. (1986), A theorem on average Ljapunov functions, *Monatsh. Math.*, **98**, 267–275.

Hutson, V. and Law, R. (1985), Permanent coexistence in general models of three interacting species, *J. Math. Biol.*, **21**, 289–298.

Hutson, V. and Moran, W. (1982), Persistence of species obeying difference equations, 3, *Math. Biol.*, **15**, 203–213.

Hutson, V. and Vickers, G.T. (1983), A criterion for permanent co-existence of species, with an application to a two prey–one predator system, *Math. Biosciences*, **63**, 253–269.

Jansen, W. (1986), *A Permanence Theorem on Replicator Systems* (to appear).

Kirlinger, G. (1986), Permanence in Lotka–Volterra equations: linked predator prey systems (to appear).

Schuster, P. and Sigmund, K. (1983), Replicator dynamics, *J. Math. Biol.*, **100**, 533–538.

Schuster, P. and Sigmund, K. (1984), Permanence and uninvadability for deterministic population models, in P. Schuster (Ed), *Stochastic Phenomena and Chaotic Behaviour in Complex Systems*, Synergetics **21** (Springer, Berlin, Heidelberg, New York).

Schuster, P., Sigmund, K., and Wolff, R. (1979), Dynamical systems under constant organization 3: Cooperative and competitive behaviour of hypercycles, *J. Diff. Eqs.*, **32**, 357–368.

Sieveking, G. (1983), Unpublished lectures on dynamical systems.

Sigmund, K. (1985), A survey on replicator equations, in J. Casti and A. Karlquist (Eds), *Complexity, Language and Life: Mathematical Approaches*, Biomathematics **16** (Springer, Berlin, Heidelberg, New York).

Svirezev, Y.M. and Logofet, D.D. (1983), *Stability of Biological Communities* (Mir Publishers, Moscow).

Taylor, P. and Jonker, L. (1978), Evolutionarily stable strategies and game dynamics, *Math. Bioscience*, **40**, 145–156.

IV. CONTROLLED DYNAMICAL SYSTEMS

State Estimation for Dynamical Systems by Means of Ellipsoids

F. L. Chernousko
Institute for Problems in Mechanics, USSR Academy of Sciences, Moscow, USSR

1. Introduction

This paper is devoted to guaranteed estimation of attainability and uncertainty sets for controlled dynamical systems. It contains a survey of results obtained recently by the author and his colleagues at the Institute for Problems in Mechanics, USSR Academy of Sciences. These results have been partly published in papers: Chernousko (1980a, 1980b, 1981, 1982, 1983), Chernousko *et al.* (1983), Ovseevich (1983), and Ovseevich and Chernousko (1982).

The method developed in these papers may be called the method of ellipsoids because it approximates the attainable sets by means of ellipsoids. Two-sided ellipsoidal estimates which are optimal with respect to volume are obtained. These estimates are connected with two specific nonlinear systems of ordinary differential equations which describe the evolution of ellipsoids that approximate attainable sets.

First (Section 2) we discuss the role of attainable sets in control theory and state the problem of approximation of these sets by means of ellipsoids. Then (Section 3) we describe algebraic operations with ellipsoids that are essential for the method developed below. In Section 4 we give the basic result of the method of ellipsoids, namely, the differential equations that describe the evolution of the approximating ellipsoids. Later (Section 5) we present some properties of these nonlinear systems of differential equations including the canonical forms of equations, their asymptotic behavior, etc. Section 6 is devoted to some generalizations of the method, especially for nonlinear systems. In Section 7 we discuss some applications of the method of ellipsoids to different control problems including optimal control and differential games. The last section (Section 8) is devoted to the application of the method to problems of guaranteed filtering in the presence of observations subject to noises.

2. Attainable Sets

We consider a controlled dynamical system described by the following differential equation, constraint, and initial condition:

$$\dot{x} = f(x, u, t) \qquad u(t) \in U[t, x(t)]$$
$$x(s) \in M, \qquad t \geq s \tag{1}$$

Here t is time, $x(t) \in R^n$ is a state vector, $u(t) \in R^m$ is a vector of control or disturbance subject to a constraint, $U(t, x)$ is a closed set in R^m which can depend on t and x, s is the initial time moment, and M is a closed initial set in R^n. If we denote by $X(x, t)$ a set of all admissible values of f in equation (1), i.e.,

$$X(x, t) = f[x, U(t, x), t] \tag{2}$$

then the system (1) can be replaced by the following differential inclusion:

$$\dot{x} \in X(x, t) \qquad x(s) \in M \tag{3}$$

The attainable set $D(t, s, M)$ for a system (1) or (2)–(3) for $t \geq s$ is a set of all vectors $x(t)$ which are values of all functions $x(\tau)$ that satisfy conditions (1) or (3) for $\tau \in [s, t]$. The attainable sets have the following important evolutional or semigroup property:

$$D(t, s, M) = D[t, \tau, D(\tau, s, M)] \tag{4}$$

for all $\tau \in [s, t]$.

Attainable sets play an important role in control theory. Solutions of many problems can be expressed in terms of these sets. Attainable sets can be used to evaluate controllability properties [if u in system (1) is a control], for estimation of perturbations and practical stability analysis [if u in system (1) is an unknown but bounded disturbance], for solution of optimal control problems and differential games, and for guaranteed filtering in dynamical systems in the presence of bounded disturbances and observation noises.

The properties of attainable sets were studied by Krasovsky (1968), Lee and Markus (1967), and other authors. Krasovsky (1968, 1970) widely applied attainable sets in optimal control theory, differential games, and guaranteed observation. Kurzhanski (1977) developed some methods of guaranteed filtering based on operations with attainable and informational sets. Formalsky (1974) studied attainable sets in the presence of integral constraints. Papers by Schweppe (1968, 1972), Schlaepfer and Schweppe (1972), and Bertsekas and Rhodes (1971) were dedicated to specific ellipsoidal approximations of sets in the state space. Some properties of attainable sets were studied by Dontchev and Veliov (1983) and Veliov (1984). The number of papers devoted to this subject is very large; here we have mentioned only some of them.

There are several possible approaches to obtaining attainable sets. One of them consists in polyhedral approximations of attainable sets and is closely connected to the description of these sets by means of support

functions. This approach can in principle produce accurate approximations but it requires a lot of computation, especially when the dimension n of the state space is large.

The other approach consists in approximating attainable sets by means of some canonical shapes described by a limited number of parameters. Here the approximation error is finite and cannot be made very small; however, this approach makes it possible to deal with sets of a fixed class, which is convenient for applications.

Such classes of sets as parallelepipeds and simplexes in R^n are described by $n + n^2$ parameters; the classes of ellipsoids and rectangular parallelepipeds need fewer [only $n + n(n + 1) / 2$] parameters. However, rectangular parallelepipeds are not invariant with respect to linear transformations. The class of ellipsoids has some other useful properties: it is closely connected with quadratic forms and Gaussian probability distributions. The possibilities of approximation of convex sets by means of ellipsoids are illustrated by the following well-known geometrical results due to Leichtweiss and F. John. For any convex set K in R^n there exists an ellipsoid E such that $E \subset K \subset nE$.

If K has a center of symmetry, then $E \subset K \subset \sqrt{n}\ E$. Here kE is an ellipsoid homothetic to E with axes k times longer than those of E.

We shall deal below with ellipsoidal approximations of attainable sets. Consider a linear controlled system

$$\dot{x} = C(t)x + u \tag{5}$$

and its finite-difference approximation

$$x(t + h) \approx [I + hC(t)]\, x(t) + hu(t) \tag{6}$$

In equations (5) and (6), $C(t)$ is a given matrix, h is a small positive number, I denotes everywhere a unity matrix. If $x(t)$ and $u(t)$ in equation (6) belong to some ellipsoids, then it can be seen from equation (6) that, in order to obtain an ellipsoid for $x(t + h)$, the following operations with ellipsoids are essential: linear transformation of an ellipsoid and summing of two ellipsoids. The third operation, namely the intersection of two ellipsoids, arises in guaranteed filtering.

3. Basic Algebraic Operations

3.1. Notation

We denote by $E(a, Q)$ an ellipsoid in R^n defined by the inequality

$$E(a, Q) = \{x: (x - a)^{\mathrm{T}} Q^{-1} (x - a) \leq 1\} \tag{7}$$

Here a is an n-vector of the center of an ellipsoid, Q is a symmetrical positive-definite $n \times n$-matrix. If Q tends to zero ($Q \to 0$), then $E(a, Q)$ degenerates into a point $x = a$.

3.2. Linear transformation

If $x \in E(a, Q)$, then

$$Ax + b \in E\,(Aa + b, A\,Q\,A^{\mathrm{T}}) \tag{8}$$

Here A is a nondegenerate $n \times n$-matrix, and b is an n-vector.

3.3. Sum of ellipsoids

The sum S of two ellipsoids is a set defined by

$$\begin{aligned} &x = x_1 + x_2 \in S \\ &x_1 \in E(a_1, Q_1) \\ &x_2 \in E(a_2, Q_2) \end{aligned} \tag{9}$$

The convex set S from equation (9) is not an ellipsoid in the general case. We shall approximate S by means of external and internal ellipsoids optimal with respect to their volume. These ellipsoids satisfy the following inclusions and optimality conditions:

$$\begin{aligned} &E(a^-, Q^-) \subset S \subset E(a^+, Q^+) \\ &\det Q^- \to \max \qquad \det Q^+ \to \min \end{aligned} \tag{10}$$

The solutions of optimization problems (10) were obtained in the explicit form:

$$\begin{aligned} &a^+ = a^- = a_1 + a_2 \\ &Q^+ = (p^{-1} + 1)\,Q_1 + (p + 1)\,Q_2 \\ &Q^- = A^{-1}\,D\,(A^{-1})^{\mathrm{T}}, \qquad D = (D_1^{0.5} + D_2^{0.5})^2 \end{aligned} \tag{11}$$

Here $p > 0$ is a unique positive root of the algebraic equation

$$\sum_{j=1}^{n} \frac{1}{\lambda_j + p} = \frac{n}{p\,(p + 1)} \tag{12}$$

where $\lambda_j > 0$, $j = 1, \ldots, n$, are eigenvalues that satisfy the equation

$$\det (Q_1 - \lambda Q_2) = 0 \tag{13}$$

The nondegenerate $n \times n$-matrix A in equation (11) is defined by the condition that it transforms both positive-definite matrices Q_1, Q_2 to diagonal matrices D_1, D_2:

$$D_i = A\, Q_i\, A^{\mathrm{T}}, \qquad i = 1, 2 \tag{14}$$

Formulas (11)–(14) describe both internal and external optimal ellipsoidal approximations for the sum of two ellipsoids.

3.4. Intersection of ellipsoids

We are interested in the optimal external ellipsoidal approximation $E(a, Q)$ for the intersection of two ellipsoids

$$P = E(a_1, Q_1) \cap E(a_2, Q_2) \subset E(a, Q) \tag{15}$$
$$\det Q \to \min$$

The problem (15), however, cannot be solved in the explicit form for the general case. Only solutions for some particular cases are known when one of the ellipsoids degenerates into a half-space or into a layer (see Shor and Gershovich, 1979). We developed some suboptimal operations of intersection which are optimal in particular cases and give a satisfactory approximation in the general case. Three versions of this operation are of practical use.

4. Equations of Ellipsoids

We consider now the linear system with ellipsoidal constraints

$$\begin{aligned} &\dot{x} = C(t)x + f(t) + u \qquad u \in E[0, G(t)] \\ &x(s) \in E(a_0, Q_0), \qquad t \geq s \end{aligned} \tag{16}$$

Here $C(t)$, $G(t)$, and Q_0 are $n \times n$-matrices, $f(t)$ and a_0 are n-vectors, and matrices $G(t)$ and Q_0 are symmetrical and positive-definite. We wish to obtain two-sided ellipsoidal estimates of the attainable sets $D(t, s, M)$ for the system (16):

$$E[a^-(t), Q^-(t)] \subset D(t, s, M) \subset E[a^+(t), Q^+(t)] \tag{17}$$

for $t \geq s$. We impose the following evolutional properties on the ellipsoids in the inclusion (17) [compare with equation (4)]:

$$\begin{aligned} &E[a^-(t), Q^-(t)] \subset D\{t, \tau, E[a^-(\tau), Q^-(\tau)]\} \\ &E[a^+(t), Q^+(t)] \supset D\{t, \tau, E[a^+(\tau), Q^+(\tau)]\} \end{aligned} \tag{18}$$

for all $t \geq \tau \geq s$. Besides that, we require that the rate of volumes v^-, v^+ of the ellipsoids in the inclusion (17) be optimal, i.e.,

$$dv^-/dt \to \max \qquad dv^+/dt \to \min \tag{19}$$

It was shown that ellipsoids that satisfy the conditions (18) and (19) exist and are unique, and that their parameters satisfy the following differential equations and initial conditions:

$$\begin{aligned} &a^+ = a^- = a(t) \\ &\dot{a} = C(t)a + f(t), \qquad a(s) = a_0 \\ &\dot{Q}^- = CQ^- + Q^- C^{\mathrm{T}} + 2G^{0.5}(G^{-0.5}Q^-G^{-0.5})^{0.5} G^{0.5} \\ &\dot{Q}^+ = CQ^+ + Q^+C^{\mathrm{T}} + hQ^+ + h^{-1}G \\ &h = \{n^{-1}\mathrm{Tr}\,[(Q^+)^{-1} G]\}^{0.5} \\ &Q^-(s) = Q^+(s) = Q_0 \end{aligned} \tag{20}$$

Note that the equation for the common center $a(t)$ of both ellipsoids is linear while the matrices Q^-, Q^+ satisfy nonlinear equations. Integrating equations (20) we can obtain the desired two-sided estimates for any time moment.

5. Properties of Ellipsoids

Here we describe briefly some properties of ellipsoids obtained from the analysis of equations (20). By means of a transformation

$$Q^{\pm} = V Z^{\pm} V^{\mathrm{T}} \tag{21}$$

where $V(t)$ is a nondegenerate matrix and $Z^{\pm}$ are new variables, equations (20) can be simplified.

If we choose $V(t)$ to be a fundamental matrix of the original system (16),

$$\dot{V} = C(t)V, \qquad V(s) = I \tag{22}$$

then the equations for $Z^{\pm}$ take the same form (20) where C is replaced by zero.

On the other hand, if we take $V = G^{0.5}$ in equation (21) then equations (20) for $Z^{\pm}$ contain a unity matrix instead of G. Therefore we can always put either $C = 0$ or $G = I$ in equations (20) for Q^{-}, Q^{+} without loss of generality.

When the internal and external ellipsoids coincide for all $t \geq s$, is the attainable set then an ellipsoid? The answer is: $a^{-} = a^{+}$ and $Q^{-} = Q^{+}$ for all $t \geq s$ if and only if the following equalities hold:

$$\begin{aligned} &CG + GC^{\mathrm{T}} - \dot{G} = \mu(t)G, \qquad t \geq s \\ &Q_0 = \lambda_0 G(s) \qquad \lambda_0 > 0 \end{aligned} \tag{23}$$

Here $\mu(t)$ is an arbitrary scalar function, and $\lambda_0 > 0$ is a positive number.

The important particular case arises if the initial ellipsoid degenerates into a point: $Q_0 = 0$. Equations (20) have singularities for $Q = 0$; hence it is necessary to analyze the asymptotic behavior of ellipsoids for $Q_0 = 0$. Without loss of generality we assume $G = I$. Let C have an expansion near the initial point

$$C = A_0 + A_1\vartheta + O(\vartheta^2) \qquad \vartheta = t - s \geq 0 \tag{24}$$

Then the solutions Q^{-}, Q^{+} have expansions

$$\begin{aligned} Q^{-}(t) &= \vartheta^2 I + \vartheta^3 D_0 + \vartheta^4 D^{-} + \cdots \\ Q^{+}(t) &= \vartheta^2 I + \vartheta^3 D_0 + \vartheta^4 D^{+} + \cdots \end{aligned} \tag{25}$$

The coefficients in equation (25) are given by

$$\begin{aligned} &D_0 = (A_0 + A_0^{\mathrm{T}})/2 \qquad D_1 = (A_1 + A_1^{\mathrm{T}})/2 \\ &D^{-} = (7/12)D_0^2 + (2/3)D_1 \\ &D^{+} = (2/3)(D_0^2 + D_1) + (1/12)n^{-2}(\mathrm{Tr}\, D_0)^2 I - (1/6)n^{-1}(\mathrm{Tr}\, D_0)D_0 \end{aligned} \tag{26}$$

The asymptotic expansions (25) and (26) are useful for starting numerical integration of equations (20) with zero initial conditions. The more complicated case when the matrix Q_0 of the initial ellipsoid is degenerate but nonzero was also studied.

Asymptotic behavior of ellipsoids when $t \to \infty$ was investigated in a particular case of a constant diagonal matrix C and diagonal solutions $Q^-(t)$, $Q^+(t)$. The results obtained were compared with the asymptotic behavior of attainable sets for $t \to \infty$.

Some *a priori* estimates for the volumes v^-, v^+ of ellipsoids (17) were obtained. We present here only one such estimate, namely

$$\frac{v^+(t)}{v_0} \le \left[\frac{v^-(t)}{v_0}\right]^{\sqrt{n}}, \qquad t \ge s \tag{27}$$

Here we assume $C = 0$ and denote $v_0 = v^-(s) = v^+(s)$. The inequality (27) together with the evident one $v^- \le v^+$ makes it possible to obtain two-sided estimates for the volume of one of the ellipsoids (internal or external) through the volume of the other one.

6. Nonlinear Systems

The results presented above can be generalized in different directions. We consider again a general nonlinear system (1) or (3) and assume that the following inclusions are true:

$$E[C^-(t)x + f^-(t), G^-(t)] \subset X(x, t) \subset E[C^+(t)x + f^+(t), G^+(t)]$$
$$E(a_0^-, Q_0^-) \subset M \subset E(a_0^+, Q_0^+) \tag{28}$$

The inclusions (28) mean that there exist two linear systems of the type (16) that bound the nonlinear system (1) or (3). If conditions (28) are satisfied, then the two-sided estimates (17) are true for attainable sets of the nonlinear system. Vectors a^-, a^+ and matrices Q^-, Q^+ in the inclusions (17) satisfy equations (20) in which C, f, G, a_0, and Q_0 must be replaced by C^-, f^-, G^-, a_0^-, and Q_0^- for a^-, Q^- and by C^+, f^+, G^+, a_0^+, and Q_0^+ for a^+, Q^+.

Most of the results described above can be applied also to systems with discrete time (to multistage processes). In this case differential equations are replaced by finite-difference ones.

More complicated approximations of sets by means of ellipsoids can be used. For instance, we can approximate the initial set M by means of the intersection of two (or more) ellipsoids

$$M \subset \left[E(a_0^1, Q_0^1) \cap E(a_0^2, Q_0^2)\right] \tag{29}$$

Then we integrate equations (20) for two initial sets that correspond to the ellipsoids (29). At each time moment the attainable set belongs to the intersection of two ellipsoids obtained by this integration. This approach

requires more computation but can give better accuracy of approximation, especially for nonsmooth initial sets, such as a rectangle.

7. Applications

The method of ellipsoids described above has different applications to control problems.

If $u(t)$ in system (1) is some unknown disturbance bound by the set U, then the external ellipsoid (17) gives the guaranteed outer bound for all possible motions for all $t \geq s$.

On the other hand, if $u(t)$ in system (1) is a control, then the internal ellipsoid indicates the inner bound for the attainable set. This ellipsoid can be used for constructing control $u(t)$ that leads to any given point belonging to this internal ellipsoid at a given time moment.

Ellipsoidal approximations can be used for obtaining two-sided estimates of a functional in optimal control. Consider the problem of minimizing the functional

$$J = F[x(T)] \to \min \qquad T > s \tag{30}$$

for the system (1); here T is a fixed terminal time and $F(x)$ is a given scalar function. The minimal value of the functional (30) is

$$J^* = \min_{x \in D(T, s, M)} F(x) \tag{31}$$

The following estimates are obvious:

$$\begin{aligned} &\min_{x \in E^+} F(x) \leq J^* \leq \min_{x \in E^-} F(x) \\ &E^{\pm} = E[a^{\pm}(T), Q^{\pm}(T)] \end{aligned} \tag{32}$$

Hence, obtaining ellipsoids $E^{\pm}$ and solving the nonlinear programming problems (32), we obtain two-sided estimates for J^* in equation (31). If the function $F(x)$ in (30) changes, then we have to solve only new nonlinear programming problems: the ellipsoids do not change in this case. Similar estimates can be obtained also for time-optimal, multicriterial, and other kinds of optimization problems.

Estimates in differential games can be obtained by means of Krasovsky's extremal rule. This rule says that in the so-called regular case the time t^* of pursuit is determined as a first time moment when the attainable set of a pursuer contains the attainable set of an evader, i.e.,

$$D_p(t^*) \supset D_e(t^*) \tag{33}$$

Two-sided estimates for t^* are given by $t_1 \le t^* \le t_2$ where t_1, t_2 are defined by inclusions similar to the inclusion (33) but written for the appropriate external and internal ellipsoids for the pursuer and the evader:

$$\begin{aligned} &t_1 : E_p^+(t_1) \supset E_e^-(t_1) \\ &t_2 : E_p^-(t_2) \supset E_e^+(t_2) \end{aligned} \tag{34}$$

8. Guaranteed Filtering

We consider the system (1) or (3) and the equation of observation

$$\begin{aligned} &y(t) = H(t)x(t) + v(t) \\ &v(t) \in E[0, B(t)] \end{aligned} \tag{35}$$

Here H is a given $r \times n$-matrix, $y(t)$ is an r-vector of observation results, $v(t)$ is a vector of observation errors bound by an ellipsoidal constraint, and $B(t)$ is a symmetrical positive-definite $r \times r$-matrix. Having obtained the observation results $y(\tau)$ for $\tau \in [s, t]$ we can estimate the set $P(t)$ to which the state vector $x(t)$ belongs; again, ellipsoidal estimates are considered:

$$x(t) \in P(t) \subset E[a(t), Q(t)] \tag{36}$$

The problem of guaranteed ellipsoidal filtering consists in obtaining parameters $a(t)$, $Q(t)$ of an ellipsoid (36). This problem was considered both for discrete observations when $y(t)$ in equation (35) is available at $t = t_0, t_1, \ldots$ and for continuous observations when equation (35) holds for all $t \ge s$.

In the first case the parameters a, Q in the inclusion (36) satisfy differential equations (20) for a^+, Q^+ between observation times. At observation times, $a(t)$ and $Q(t)$ are discontinuous; they have jumps corresponding to the immediate change of ellipsoids due to observations. Different formulas for intersection of ellipsoids can be used here for calculating these jumps.

The continuous filtering is governed by a specific system of differential equations which takes into account the dynamics of the system and results of observations.

The algorithms of state estimation based on the method of ellipsoids described above were realized in a package of FORTRAN programs. These programs integrate the equations of internal and external ellipsoids and simulate both discrete and continuous guaranteed filtering for systems with $n \le 10$.

As a result of computer simulation a number of numerical data are obtained, including approximation of attainable sets of different stable and unstable systems, estimates in optimal control and differential games, use of

intersection of ellipsoids for nonsmooth initial sets, simulation of discrete-time and continuous-time filtering, etc. Part of these numerical results as well as proofs of the theorems and additional details are presented in the papers mentioned in the Introduction.

References

Bertsekas, D. P. and Rhodes, I. B. (1971), Recursive state estimation for a set-membership description of uncertainty, *IEEE Transactions on Automatic Control*, **AC-16**, 117–128.

Chernousko, F. L. (1980a), Guaranteed estimates of undetermined quantities by means of ellipsoids, *Soviet Mathematics Doklady*, **251**(1), 51–54.

Chernousko, F. L. (1980b), Optimal guaranteed estimates of uncertainties by means of ellipsoids, Parts I, II, III, *Izvestiya of the USSR Academy of Sciences, Engineering Cybernetics*, **3**, 3–11; **4**, 3–11; **5**, 5–11.

Chernousko, F. L. (1981), Ellipsoidal estimates for attainable set of a controlled system, *Applied Mathematics and Mechanics*, **45**(1), 11–19.

Chernousko, F. L. (1982), Ellipsoidal bounds for sets of attainability and uncertainty in control problems, *Optimal Control Applications and Methods*, **3**(2), 187–202.

Chernousko, F. L. (1983), On equations of ellipsoids approximating reachable sets, *Problems of Control and Information Theory*, **12**(2), 97–110.

Chernousko, F. L., Ovseevich, A. I., Klepfish, B. R., and Trushchenkov, V. L. (1983), *Ellipsoidal Estimation of State for Controlled Dynamic Systems*, Preprint No. 224 (Institute for Problems in Mechanics, USSR Academy of Sciences, Moscow).

Dontchev, A. L. and Veliov, V. M. (1983), On the behavior of solutions of linear autonomous differential inclusions at infinity, *Comptes rendus de l'Academie Bulgare des Sciences*, **36**(8), 1021–1024.

Formalsky, A. M. (1974), *Controllability and Stability of Systems with Restricted Resources* (Nauka, Moscow).

Krasovsky, N. N. (1968), *Theory of Control of Motion* (Nauka, Moscow).

Krasovsky, N. N. (1970), *Game Problems of Meeting of Motions* (Nauka, Moscow).

Kurzhanski, A. B. (1977), *Control and Observation in Conditions of Uncertainty* (Nauka, Moscow).

Lee, E. B. and Markus, L. (1967), *Foundations of Optimal Control Theory* (Wiley, New York).

Ovseevich, A. I. (1983), Extremal properties of ellipsoids approximating reachable sets, *Problems of Control and Information Theory*, **12**(1), 43–54.

Ovseevich, A. J. and Chernousko, F. L. (1982). Two-sided estimates of attainable sets for controlled systems, *Applied Mathematics and Mechanics*, **46**(5), 737–744.

Schlaepfer, F. M. and Schweppe, F. C. (1972), Continuous-time state estimation under disturbances bounded by convex sets, *IEEE Transactions on Automatic Control*, **AC-17**, 197–205.

Schweppe, F. C. (1968), Recursive state estimation: unknown but bounded errors and system inputs, *IEEE Transactions on Automatic Control*, **AC–13**, 22–28.

Schweppe, F. C. (1972), *Uncertain Dynamic Systems* (Prentice Hall, Englewood Cliffs, NJ).

Shor, N. Z. and Gershovich, V. I. (1979), On one family of algorithms for the solution of convex programming problems, *Kibernetika*, **4**, 62–67.

Veliov, V. M. (1984), On the local properties of Bellman's function for nonlinear time-optimal control problems, *Serdica-Bulgariae Mathematicae Publicationes*, **10**, 68–77.

Singularity Theory for Nonlinear Optimization Problems

J. Casti
International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

1. Background

Consider a smooth (C^∞) function $f: R^n \to R^m$ and assume that f has a critical point at the origin, i.e., $df(0) = 0$. The theory of singularities as developed by Thom, Mather, Arnol'd, and others (Lu, 1976; Gibson, 1979; Arnol'd, 1981) addresses the following basic questions:

(1) What is the local character of f in a neighborhood of the critical point? Basically, this question amounts to asking "at what point is it safe to truncate the Taylor series for f?" This is the *determinacy* problem.
(2) What are the "essential" perturbations of f? That is, what perturbations of f can occur that change the qualitative nature of f and that cannot be transformed away by a change of coordinates? This is the *unfolding* problem.
(3) Can we classify the types of singularities that f can have up to diffeomorphism? This is the *classification* problem.

Elementary catastrophe theory largely solves these three problems (when $m = 1$); its generalization to singularity theory solves the first two, and gives relatively complete information on the third for m, n small. Here we outline a program for the utilization of these results in an applied setting to deal with certain types of nonlinear optimization problems. In the following section we give a brief summary of the main results of singularity theory for problems (1)–(3) for *functions* ($m = 1$) and then proceed to a discussion of how these results may be employed for nonlinear optimization.

2. Determinacy, Unfoldings, and Classifications

2.1. Equivalence of germs

In its local version, elementary catastrophe theory deals with functions $f: U \to R$ where U is a neighborhood of O in R^n. The cleanest way to handle

such functions is to pass to *germs*, a germ being a class of functions that agrees on suitable neighborhoods of O. All operations on germs are defined by performing similar operations on representatives of their classes. In the sequel, we usually make no distinction between a germ and a representative function.

We let E_n be the set of all smooth germs $R^n \to R$, and let E_{nm} be the set of all smooth germs $R^n \to R^m$. Of course $E_{n,1} = E_n$. These sets are vector spaces over R, of infinite dimension. We abbreviate $(x_1, \ldots, x_n) \in R^n$ to x. If $f \in E_{nm}$ then

$$f(x) = [f_1(x), \ldots, f_m(x)]$$

and the f_i are the *components* of f.

A *diffeomorphism* germ $\varphi : R^n \to R^n$ satisfies $\varphi(0) = 0$, and has an inverse φ' such that $\varphi[\varphi'(x)] = x = \varphi'[\varphi(x)]$ for x near 0. It represents a smooth, invertible local coordinate change. By the Inverse Function Theorem, φ is a diffeomorphism germ if and only if it has a nonzero Jacobian, that is,

$$\det[\partial \varphi_i / \partial x_j (0)] \neq 0$$

Two germs $f, g : R^n \to R$ are *right equivalent* if there is a diffeomorphism germ φ and a constant $\gamma \in R$ such that

$$g(x) = f\varphi(x) + \gamma$$

This is the natural equivalence for studying topological properties of the gradient ∇f (Poston and Stewart, 1978). If f, rather than ∇f, is important, the term γ is omitted.

A *type* of germ is a right equivalence class and the classification of germs up to right equivalence amounts to a classification of types. Each type forms a subset of E_n, and the central object of study is the way these types fit together.

A precise description is complicated by the fact that most types have infinite dimension; but there is a measure of the complexity of a type, the *codimension*, which is generally finite. Heuristically, it is the difference between the dimension of the type and that of E_n (even though both are infinite). A precise definition is given below.

The largest types have codimension 0 and form open sets in E_n. Their boundaries contain types of codimension 1; the boundaries of these in turn contain types of codimension 2, and so on, with higher codimensions revealing progressively more complex types. Types of infinite codimension exist, but form a very small set in a reasonable sense.

2.2. Codimension and the Jacobian ideal

Let E_n be the set of germs $R^n \to R$, and let F be the set of formal power series in $x_1, \ldots, x_n$. There is a map $j: E \to F$ defined by

$$jf = f(0) + \sum \frac{\partial f}{\partial x_i}(0)x_i + 0.5 \sum \frac{\partial^2}{\partial x_i \partial x_j}(0)x_i x_j + \cdots$$

where the right-hand side is the Taylor series, or *jet*, of f. Note that it exists as a *formal* power series for all smooth f: convergence is not required in what follows. The map j is onto, linear over R, and preserves products [i.e., $j(f \cdot g) = j(fg) = (jf \cdot jg)$].

Let m_n be the set of $f \in E_n$ such that $f(0) = 0$. This is an *ideal* of E_n (meaning that if $f \in m_n$ and $g \in E_n$ then $fg \in m_n$, which we write briefly as $m_n E_n \subseteq m_n$). Its kth power m_n^k consists of all $f \in E_n$ such that

$$0 = f(0) = df(0) = d^2 f(0) = \cdots = d^{k-1} f(0)$$

In particular, f is a *singularity* if and only if $f \in m_n$. The ideals m_n^k form a decreasing sequence.

$$E_n \supset m_n \supset m_n^2 \supset m_n^3 \supset \cdots$$

There is a similar chain in F_n. Let $M_n = j(m_n)$: this is the set of formal power series with zero constant term. Then $M_n^k = j(m_n^k)$ is the set of formal power series without terms of degree $\leq k-1$. The intersection of all M_n^k is 0; the intersection of all m_n^k is the set m_n^∞ of flat germs, having zero Taylor series.

The *Jacobian ideal* of a singularity f is the set of all germs expressible in the form

$$g_1 \frac{\partial f}{\partial x_1} + \cdots + g_n \frac{\partial f}{\partial x_n}$$

for arbitrary germs g_i. We denote it by $\Delta(f)$, or merely Δ when f is understood. Its image $j\Delta(f) \subseteq F_n$ has an analogous definition, where the partial derivatives are defined formally. Since f is a singularity, $\Delta(f) \subseteq m_n$. The *codimension* of f is defined to be

$$\operatorname{cod}(f) = \dim_R m_n / \Delta(f)$$

Similarly, at the formal power series level, we define

$$\mathrm{cod}(jf) = \dim_R M_n / j\Delta(f)$$

The codimension of an orbit is the same as that of its tangent space T. This is the same as the dimension of the quotient vector space E/T. In E_n, the analog of this tangent space is the Jacobian ideal, so the codimension should be $\dim E_n / \Delta(f)$. This measures the number of independent directions in E_n "missing" from $\Delta(f)$, or equivalently missing from the orbit of f.

The computation of $\mathrm{cod}(f)$ is effected by means of the following result: if either $\mathrm{cod}(f)$ or $\mathrm{cod}(jf)$ is finite then so is the other, and they are equal. Thus, the computation may be carried out on the formal power series level where it is a combinatorial calculation. For examples in classical notation, see Poston and Stewart (1978).

2.3. Determinacy

Let $f \in E_n$, and define the k-jet $j^k(f)$ to be the Taylor series of f up to and including terms of order k. For example,

$$j^6[\sin(x)] = x - x^3/3! + x^5/5!$$

We say that f is *k-determinate* (or k-determined) if for any $g \in E_n$ such that $j^k g = j^k f$, it follows that g is right equivalent to f.

A germ is 1-determined if its linear part is nonzero, that is, its derivative does not vanish. So the non-1-determined germs are the singularities. If f is a singularity and $f(0) = 0$ (as we can assume) then the second derivative gives the 2-jet of f in the form

$$j^2 f(x_1, \ldots, x_n) = \sum_{i,j} H_{ij} x^i x^j$$

where the Hessian matrix

$$H = (H_{ij})(0)$$

is symmetric. It can be shown that f is 2-determined if and only if $\det(H) \neq 0$; in this case f is right equivalent to

$$\pm x_1^2 \pm \cdots \pm x_n^2 \tag{1}$$

This is a reformulation in determinacy terms of the *Morse Lemma* (Milnor, 1973). A germ equivalent to expression (1) is said to be *Morse*. Morse germs are precisely those of codimension 0. The number l of negative signs in expression (1) is the *index* of f, and f is an *l-saddle*. Morse theory (Milnor, 1973) describes the *global* properties of a function $f: X \to R$ where X is a

smooth manifold and f has only Morse singularities (see Casti, 1984).

There exist rules for computing the determinacy of a given germ: an easy necessary condition, an easy (different) sufficient condition, and a harder necessary-and-sufficient condition.

Let Δ be the Jacobian ideal of f. Then:

(1) If $m_n^k \subseteq m_n\Delta$ then f is k-determined.
(2) If f is k-determined then $m_n^{k+1} \subseteq m_n\Delta$.
(3) f is k-determined if and only if $m_n^{k+1} \subseteq m_n\Delta(f+g)$ for all $g \in m_n^{k+1}$.

There is a slightly stronger form of (1), namely:

(1') If $m_n^{k+1} \subseteq m_n^2\Delta$ then f is k-determined.

Numerous examples in Poston and Stewart (1978) and Gibson (1979) show how to compute the determinacy of a given f. For example, suppose f is in Morse form, as expression (1). Then $\Delta = \langle +2x_1, \ldots, +2x_n \rangle = m_n$ and $m_n^2 = m_n\Delta$. By (1), f is 2-determined as asserted above.

A germ is *finitely* determined if it is k-determined for some finite k. The following are equivalent:

(4) f has finite codimension.
(5) f is finitely determined.
(6) $m_n^t \subseteq \Delta$ for some t.

The solution to the Determinacy Problem is thus that it is safe (up to right equivalence) to truncate a k-determined germ at degree k of its Taylor series. For a germ such as $x^2y \in E_2$, which is not finitely determined, it is not safe to truncate higher order perturbing terms (and, indeed, $x^2y + y^t$ has a type that depends on t). Germs that are not finitely determined either arise in a context where some symmetry is acting (and should be analyzed by methods similar to those above, but which take symmetry into account – which can be done) or must be viewed with suspicion. By (4), we may summarize: "nice" germs have finite codimension.

Suppose that f is not 2-determinate, so that $\det(H) = 0$. Let the rank of the matrix H be r and call $n - r$ its *corank*. A useful result, called the *Splitting Lemma*, says that f is right equivalent to a germ of the form

$$g(x_1, \ldots, x_{n-r}) \pm x_{n-r+1}^2 \pm \cdots \pm x_n^2$$

For many purposes, the quadratic terms may be ignored. So the Splitting Lemma reduces the effective number of variables to $n - r$. A simple proof for finite dimensions is in Poston and Stewart (1978).

The determinacy calculations, and the application of the Splitting Lemma, may be carried out equally well on $j^k f$ in F_n, provided the codimension of f is finite. The formal power series setting is better for computations.

2.4. Unfoldings

An unfolding of a singularity is a "parametrized family of perturbations". The notion is useful mainly because, for finite codimension singularities, there exists a "universal unfolding", which in a sense captures all possible unfoldings.

More rigorously, let $f \in E_n$. Then an *l-parameter unfolding* of f is a germ $F \in F_{n+l}$, that is, a real-valued germ of a function $F(x_1, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_l) = F(x, \varepsilon)$, such that $F(x, 0) = f(x)$.

An unfolding F is *induced* from F if

$$F(x, \delta) = F[\rho_\delta(x), \psi(\delta)] + \gamma(\delta)$$

where

$$\delta = (\delta_1, \ldots, \delta_m) \in R^m$$
$$\rho_\delta : R^n \to R^n$$
$$\psi : R^m \to R^l$$
$$\gamma : R^l \to \mathrm{R}$$

Two unfoldings are *equivalent* if each can be induced from the other. An l-parameter unfolding is *versal* if all other unfoldings can be induced from it; *universal* if, in addition, l is as small as possible.

Suppose that f has finite codimension c. Let $u_1, \ldots, u_c$ be a basis for $m_n / \Delta(f)$. Then it is a theorem that a *universal unfolding* is given by the germ

$$F(x, \varepsilon) = f(x) + \varepsilon_1 u_1(x) + \cdots + \varepsilon_c u_c(x) \qquad \varepsilon_i \in R$$

While different choices of the u_i can be made, a universal unfolding is unique up to equivalence. The existence of universal unfoldings in finite codimension, and the method for computing them, is probably the most significant and useful result in elementary catastrophe theory. [Note that equation (2) is linear in the unfolding variables ε. This is a theorem and is *not* built into the definition of an unfolding.]

For example, if $f(x, y) = x^3 + y^4$, then a basis for $m_2 / \Delta(f)$ is $\{x, y, xy, y^2, xy^2\}$. So a universal unfolding is given by

$$F(x, y, \varepsilon) + x^3 + y^4 + \varepsilon_1 x + \varepsilon_2 y + \varepsilon_3 xy + \varepsilon_4 y^2 + \varepsilon_5 xy^2$$

The codimension of a germ f has several interpretations:

(1) The codimension of the Jacobian ideal in m_n.
(2) The number of independent directions "missing" from the orbit of f.
(3) The number of parameters in any universal unfolding of f.

In addition, if the codimension of f is c, it can be shown that any small perturbation of f has at most $c + 1$ critical points.

2.5. Classification

We sketch how these ideas may be used to classify germs of codimension at most 4.

Let $f \in E_n$. If f is not a singularity then $f(x)$ is right equivalent to x_1. If f is a singularity and its Hessian has nonzero determinant, then f is right equivalent to $\pm x_1^2 \pm \cdots \pm x_n^2$. Otherwise, $\det(H)=0$. Let $k = n - r$ be the corank of H and split f as

$$f(x) = g(x_1, \ldots, x_k) \pm x_{k+1}^2 \pm \cdots + x_n^2$$

It can be proved that the classification of possibilities for f depends only on the similar classification for g.

The Taylor series of g begins with cubic or higher terms. First, suppose that $k = 1$, and let the first nonzero jet of g be $a_t x^t$. This is t-determined, and scales to $\pm x^t$ (t even), x^t (t odd). The codimension is $t-2$, so $t = 3$, 4, 5, or 6.

Next, let $k = 2$ and let

$$j^3 g(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$$

By a linear change of variable, this cubic may be brought to the form $x^3 + xy^2$ (one real root), $x^3 - xy^2$ (three distinct real roots), x^2y (three real roots, one repeated), x^3 (three real roots, all repeated), or 0.

The forms $x^3 \pm xy^2$ are 3-determined, and of codimension 3.

The form x^2y is not 3-determined, so we consider higher terms. A series of changes of variable bring any higher order expansion to the form $x^2y + y^t$, which is t-determined and of codimension t. Only $t = 4$ is relevant to our problem here.

No higher term added to x^3 produces a codimension 4 result; and no higher term added to 0 does.

Finally, let $k \geq 3$. Then the codimension can be proved to be at least 7, so this case does not arise.

Thus, we have classified the germs of codimension ≤ 4 into the canonical forms

$$
\begin{aligned}
&x_1 \\
&\pm x_1^2 \pm \cdots \pm x_n^2 \\
&x_1^3 + (M) \\
&x_1^4 + (M) \\
&x_1^5 + (M) \\
&x_1^6 + (M) \\
&x_1^3 - x_1 x_2^2 + (N) \\
&x_1^3 + x_1 x_2^2 + (N) \\
&x_1^3 + x_2^4 + (N)
\end{aligned}
$$

where

$$(M) = \pm x_2^2 \pm \cdots \pm x_n^2 \qquad (N) = \pm x_3^2 \pm \cdots \pm x_n^2$$

The celebrated elementary catastrophes of Thom are the universal unfoldings of the singularities on this list, or its extension to higher codimensions. The universal unfolding arises when we try to classify not germs, but l-parameter families of germs. For $l \leq 4$, "almost all" such are given by universal unfoldings of germs of codimension ≤ 4.

Table 1 summarizes the list of germs and their unfoldings up to codimension 5, together with their customary name and symbol in the systematic notation of Arnol'd (1981). The terms (M) and (N) are omitted for clarity, x and y replace x_1 and x_2, and unfolding parameters are listed as (a,b,c,d,e) rather than $(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4,\varepsilon_5)$.

3. Singularity Theory and Nonlinear Programming

We consider the problem

$$\max f(x) \tag{3}$$

over all $x \in R^n$ such that

$$g(x) \leq 0 \tag{4}$$

where $f, g \in m_n$. There are at least three different aspects of this standard nonlinear optimization problem which singularity theory can shed some light upon:

Table 1 The elementary catastrophes of codimension ≤ 5. When the ± sign occurs, germs with sign (+) are called *standard*, (−) are called *dual*.[a]

Symbol	*Name*	*Germ*	*Universal unfolding*	*Co-rank*	*Codi-mension*
A_2	Fold	x^3	x^3+ax	1	1
$\pm A_3$	Cusp	$\pm x^4$	$\pm x^4+ax^2+bx$	1	2
A_4	Swallowtail	x^5	$x^5+ax^3+bx^2+cx$	1	3
$\pm A_5$	Butterfly	$\pm x^6$	$\pm x^6+ax^4+bx^3+cx^2+dx$	1	4
A_6	Wigwam	x^7	$x^7+ax^5+bx^4+cx^3+dx^2+ex$	1	5
D_4^-	Elliptic umbilic	x^3-xy^2	$x^3-xy^2+ax^2+bx+cy$	2	3
D_4^+	Hyperbolic umbilic	x^3+xy^2	$x^3+xy^2+ax^2+bx+cy$	2	3
$\pm D_5$	Parabolic umbilic	$\pm(x^2y+y^4)$	$\pm(x^2y+y^4)+ax^2+by^2+cx+dy$	2	4
D_6^-	Second elliptic umbilic	x^5-xy^2	$x^5-xy^2+ay^3+bx^2+cy^2+dx+ey$	2	5
D_6^+	Second hyperbolic umbilic	x^5+xy^2	$x^5+xy^2+ay^3+bx^2+cy^2+dx+ey$	2	5
$\pm E_6$	Symbolic umbilic	$\pm(x^3+y^4)$	$\pm(x^3+y^4)+axy^2+by^2+cxy+dx+ey$	2	5

[a]The above sketch shows how the classification problem reduces to the determinacy and unfolding problems (and is relatively easy once these are solved). In applications, the main influence of the classification is an organizing one: the determinacy and unfolding theorems play a more direct role.

(1) Reduction of dimensionality in the decision space for dual, penalty, and barrier type algorithms (Bazaraa and Shetty, 1979).
(2) Transformation of the constraint space into simpler form for primal type algorithms (Bazaraa and Shetty, 1979).
(3) Sensitivity analysis.

Let us examine each of these areas in turn.

3.1. Dimensionality reduction and the splitting lemma

If the optimization problem (3)–(4) is to be approached using one of the dual penalty or barrier algorithms of Bazaraa and Shetty (1979), the Splitting Lemma can be used to reduce the dimension of the decision vector in the surrogate objective function. For example, consider the augmented Lagrangian method, for which the surrogate objective function is

$$G(x,\alpha) = \alpha' g + f + \rho/2 \,||g||^2$$

where α is a vector of multipliers and ρ is some positive constant. The parameters α are updated according to, say, the augmented Lagrangian scheme of Hestenes.

Assume that the critical point of G is located at $x = x^*, \alpha = \alpha^*$, and that the corank of $G(x,\alpha) = r$. Then the Splitting Lemma insures that there exist coordinate transformations $x \to \hat{x}$, $\alpha \to \hat{\alpha}$ such that $G \to \hat{G}$, where

$$\hat{G}(\hat{x},\hat{\alpha}) = G_1(\hat{x}_1, \ldots, \hat{x}_r, \hat{\alpha}_1, \ldots, \hat{\alpha}_c) + M(x_{r+1}, \ldots, x_n)$$

where c = codim G while $G_1(\cdot)$ is a function $O(|x|^3)$, which is linear in $\hat{\alpha}_1, \ldots, \hat{\alpha}_c$. The function $M(\cdot)$ is a pure quadratic. The important point here is that usually $r << n$, which implies that most of the computational work is involved in minimizing the quadratic M, which can be done very efficiently by any of a number of quasi-Newton schemes. The essentially nonlinear part of the problem involves the minimization of G, which, however, involves only r variables. Often $r = 1$ or 2, even if n is very large, say, hundreds, so the computational savings can be significant.

The potential drawback to the above scheme is that in order to compute r, the corank of G, we need to know the Hessian

$$H = \left[\frac{\partial^2 G}{\partial x^2}\right]$$

at the critical point (x^*,α^*). Since it is precisely x^* which we seek, it appears at first glance that the situation is not too promising. However, this problem can be circumvented in at least two different ways:

(1) Often it can be seen that the Hessian will be of constant rank in some neighborhood D of x^*, even if we do not know x^* exactly. This situation comes about since we usually have at least some idea of the region D containing x^*. Thus, if we have an estimate of D and know that rank $H(x,\alpha)$ = constant for all $x \in D$, then we can use this information in a successive approximation scheme generating a sequence $x_n \to x^*$. The idea is to apply the Splitting Lemma to each approximate problem at the point x_n.

(2) If there is no information about the rank H, then we can appeal to the inequality

$$r(r+1)/2 \leq \text{codim } G$$

which always holds. We can take a pessimistic estimate of r which, at worst, means only that we include a few more variables in our nonlinear optimization of $G_1(\cdot)$ than might have been needed. If codim $G \leq 2$, then we can see from the inequality that $r = 1$ and there is only a single essential, nonlinear variable, *regardless* of where x^* is located. Otherwise there may be several nonlinear variables, but the number will still be severely limited by the above inequality.

An essential ingredient in making the above scheme work in practice is the ease of determining the coordinate transformations $x \to \hat{x}$, $\alpha \to \hat{\alpha}$. As noted in Section 2, the theory guarantees that such transformations exist and, moreover, that they are themselves diffeomorphisms. Thus, the coordinate changes

$$\hat{x}_i = \hat{x}_i(x_1, x_2, \ldots, x_n)$$
$$\hat{\alpha} = \hat{\alpha}_j(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

have convergent power series expansions. Consequently, since we know the original form of G and its normal form $\hat{G}$, in principle we can substitute the above expansions and match coefficients in order to determine the explicit form of the transformations. The operational implementation of this idea, however, may require a substantial amount of algebra, depending upon the exact nature of G.

3.2. Simplifying the constraint space

For nonlinear constrained optimization problems having nonlinear constraint sets, the coordinate changes discussed above can be employed to "straighten-out" the binding constraints in a neighborhood of regular points, so that primal methods for solving constrained optimization problems can be used, dealing only with *linear* side constraints. The essence of the primal methods is to start with a feasible direction along which the objective function is improving. A one-dimensional line search (interval bisection, Newton's method, etc.) is then used to solve the one-dimensional optimization problem along the improving feasible direction, constrained so that the resulting solution remains feasible (Bazaraa and Shetty, 1979).

A specific example of such a primal method is the *gradient projection* technique due to Rosen. This method generates an improving feasible direction by projecting the negative of the gradient vector of f onto the affine subspace determined by the intersection of the binding constraints, assuming the constraints are *linear*. A projection matrix P is formed from a suitable linear combination of the normal vectors of the constraint subspaces (i.e., the gradients of the binding constraints). The resulting one-dimensional optimization is then guaranteed to remain feasible as long as a suitable upper bound is observed on the line search (Bazarraa and Shetty, 1979).

In the event the constraints are nonlinear, the gradient of f is projected onto the intersection of the tangent spaces to the binding constraints, so that movement along the improving feasible direction will, in general, take the solution outside the feasible region (see *Figure 1*). This necessitates a correction move to bring the solutions back into the feasible regions after the one-dimensional search has been completed. Singularity theory appears to offer the possibility of materially improving the above procedure, as we now indicate.

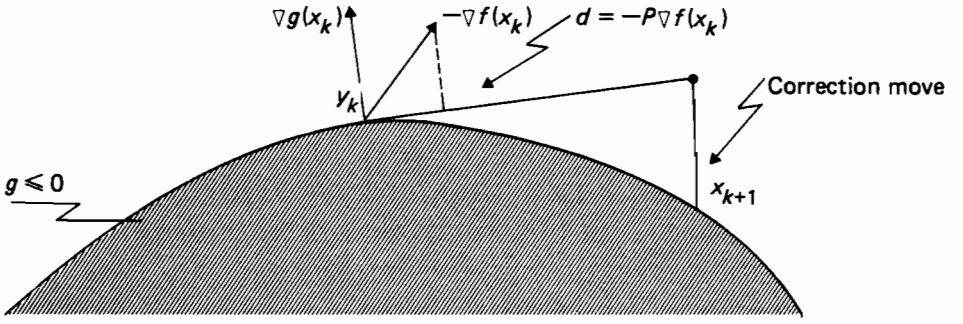


Figure 1 **Projected gradient method of Rosen for nonlinear constraints (from Figure 10.5, Bazaraa and Shetty, 1979).**

Consider the following nonlinear programming problem:

$$\begin{aligned} &\text{minimize:} && f(x) \\ &\text{subject to:} && g_i(x) \leq 0 \quad i = 1,2,\ldots,m \\ & && x \geq 0 \end{aligned}$$

For any x such that $x \geq 0$, if $I = \{i: g_i(x) = 0\}$, then

$$X = \{x: g_i(x) = 0\} = \bigcap_{i \in I} [g_i(x) \cap R^{n-1} \text{ hyperplane}]$$

will be the intersection of a finite number of manifolds in R^n and thus, with the possible exclusion of a set of points of codimension n (corners), will inherit the manifold structure locally. Locally, then, a coordinate change could be effected in X which will cause X to take the form

$$X \to Y = \{y: 0 = a'y + c, \ a,c \text{ constant vectors}\}$$

as long as the gradients of the binding constraints do not vanish. A transversality argument can be used to rule out the latter possibility.

Assuming that only the constraint $g_i(x) = 0$ is binding, let

$$S_i = T_x\, g_i(x) \cap R^{n-1} \text{ hyperplane}$$

where

$$T_x g_i(x) = \text{tangent space to } g_i \text{ at } x$$

Since codim $T_x g_i(x) = 1$ and codim{R^{n-1} hyperplane} = 1, if the intersection is transverse

$$\text{codim } T_x g_i(x) + \text{codim}\{R^{n-1} \text{ hyperplane}\} = \text{codim } S_i = 2$$

Results from differential topology assert that the set of critical points R_i for g_i will be isolated; thus the dim $R_i = 0$ and codim $R_i = n$. Therefore,

$$\text{codim } R_i + \text{codim } S_i = n + 2 > n$$

So, for $\nabla g_i(x)$ to be zero at exactly the same points where $g_i(x) = 0$ constitutes a nontransverse intersection and is therefore nongeneric. If any such points should occur, they will be isolated and thus not form a constraint boundary.

In practice, finding X and the coordinate transformation necessary to make it look like Y usually requires some effort. However, if projection onto only one binding constraint is necessary, the calculation becomes simpler, as the following example shows:

$$\min f(x_1, x_2) = 0.5\, x_1^2 + 0.5\, x_2^2 - x_1 - x_2$$

(the geometry in x space is shown in *Figure 2*) subject to

$$\begin{aligned} x_1^2 + x_2^2 - 1 &\leq 0 \\ -x_1 &\leq 0 \\ -x_2 &\leq 0 \end{aligned}$$

$\nabla f(x) = (x_1 - 1, x_2 - 1)$	at (1,0):	$\nabla f(1,0) = (0,-1)$	
$\nabla g_1(x) = (2x_1, 2x_2)$		$\nabla g_1(1,0) = (2,0)$	binding
$\nabla g_2(x) = (-1,0)$		$\nabla g_2(1,0) = (-1,0)$	
$\nabla g_3(x) = (0,-1)$		$\nabla g_3(1,0) = (0,-1)$	binding

As can be seen, we want to project onto $g_1(x)$. To straighten out g_1, let $y_1 = x_1^2$, $y_2 = x_2^2$. In the new coordinates, ∇f_{new} will be

$$\nabla f_{\text{new}}(y) = (y_1^{0.5} - 1, y_2^{0.5} - 1), \qquad \nabla f_{\text{new}}(1,0) = (0,-1)$$

(Note: This is not the gradient of the transformed objective function but rather the transformed gradient of the old objective function.)

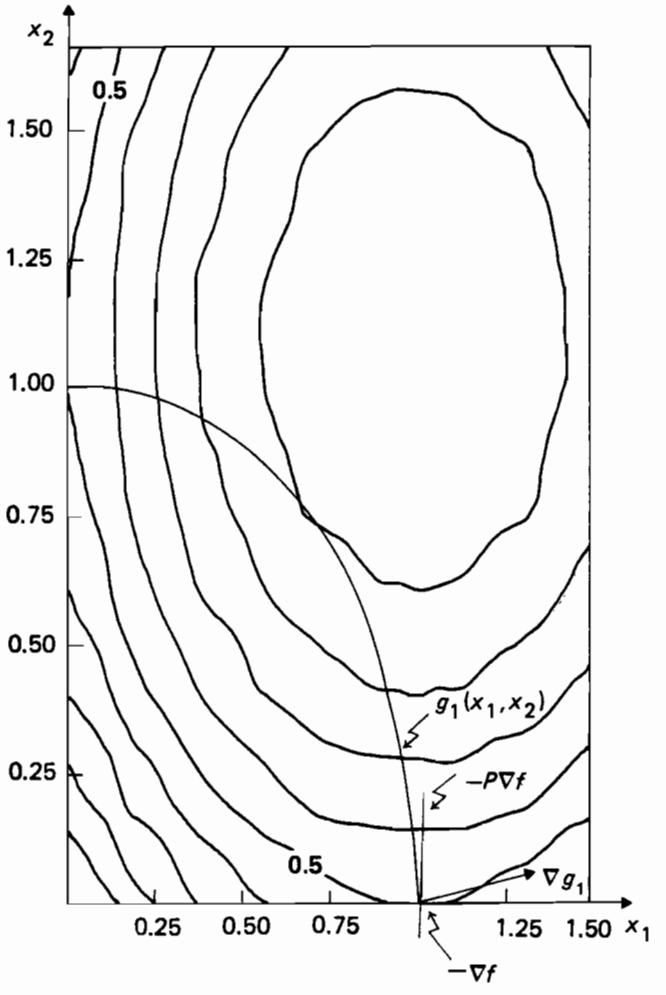


Figure 2 Configuration in x space.

The new problem is

$$\min f(y_1, y_2) = 0.5y_1 + 0.5y_2 - y_1^{0.5} - y_2^{0.5}$$

(the geometry in y space is shown in *Figure 3*) subject to

$$y_1 + y_2 - 1 \leq 0$$
$$-y_1 \leq 0$$
$$-y_2 \leq 0$$

Now the constraint is linear and we project ∇f_{new} onto g_1 by forming the projection matrix:

$$\nabla g_1 = M = (1\ 1) \qquad MM^t = (1\ 1)\begin{pmatrix}1\\1\end{pmatrix} = 2 \qquad (MM^t)^{-1} = 1/2$$

$$P = I - M^t(MM^t)^{-1}M = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix} - \begin{pmatrix}1\\1\end{pmatrix}(1/2)\,(1\ 1) = \begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}$$

$$d = -P\nabla f_{\text{new}} = -\begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}-\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}$$

The objective function is optimized along the constraint by letting

$$y = \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}-0.5\,h\\ 0.5\,h\end{pmatrix}$$

$$f(h) = 0.5(1-0.5h) + 0.25h - (1-0.5h)^{0.5} - (0.5h)^{0.5}$$

$$\frac{df}{dh} = 0.25(1-0.5h)^{-1.5} - 0.25(0.5h)^{-1.5} = 0 \Longrightarrow h = 1$$

So the minimum is taken on at

$$y = \begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix} \quad \text{or} \quad x = \begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$$

That this is the optimum can be seen by trying to form an improving feasible direction in x space. The result will be the zero vector, indicating that the optimum has been reached.

$$\nabla g_1(x) = \left(\frac{2}{\sqrt{2}}, \frac{2}{\sqrt{2}}\right) \qquad MM^t = (\sqrt{2}\ \sqrt{2})\begin{pmatrix}\sqrt{2}\\ \sqrt{2}\end{pmatrix} = 4 \qquad (MM^t)^{-1} = 0.25$$

$$\nabla f(x) = (\sqrt{2}-1,\ \sqrt{2}-1)\,\frac{1}{\sqrt{2}}$$

$$P = I - M^t(MM^t)^{-1}M = \begin{pmatrix}1 & 0\\1 & 0\end{pmatrix} - \begin{pmatrix}\sqrt{2}\\ \sqrt{2}\end{pmatrix}\frac{1}{4}\,(\sqrt{2}\ \ \sqrt{2}) = \begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}$$

$$d = P\nabla f(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}\sqrt{2}-1\\ \sqrt{2}-1\end{pmatrix} = 0$$

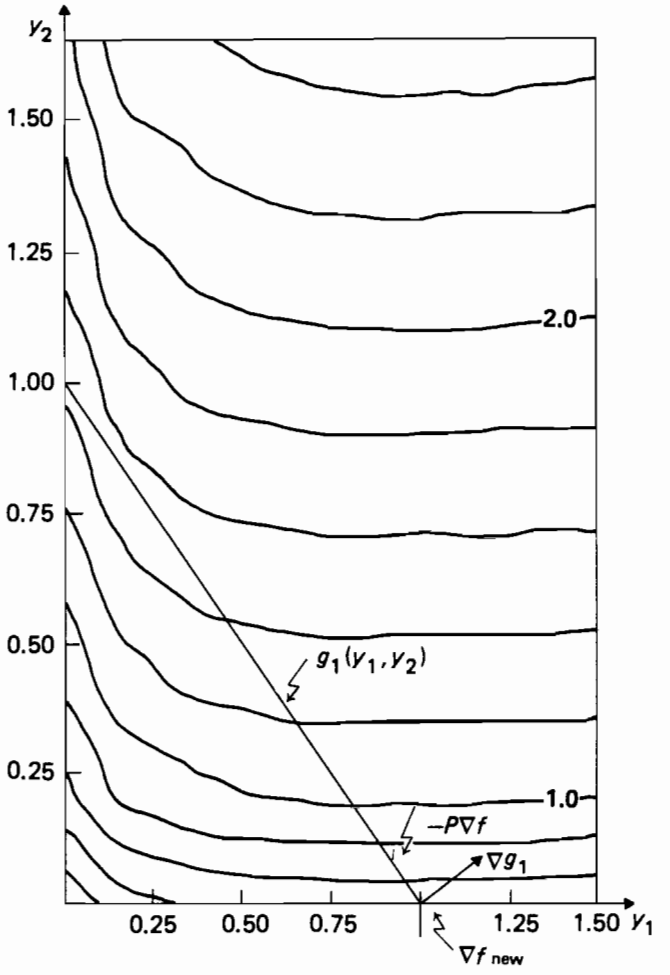


Figure 3 Configuration in y space.

as claimed. A summary of the algorithm is given next.

Initialization step: Choose a feasible point x_u and find $I_t = \{i: g_1(x) = 0\}$. Let $u = 1$ and go to *Step 1*.

Step 1: If $I_t = 0$, let $P = I$, form $d_u = P\nabla f(x_u)$ and go to *Step 3*. Otherwise, form the projection matrix in x-space as follows. Let $M = D_g(x_u)$ be the matrix of gradients of the binding constraints at x_u. If $P = I - M^t(MM^t)^{-1}M = 0$, let $W = -(MM^t)^{-1}M\nabla f(x_u)$. If $W \geq 0$, x_u will be a Kuhn–Tucker point, otherwise delete a row corresponding to $W_t \geq 0$ and repeat *Step 1*. This has the effect of eliminating binding constraints from

consideration which would not generate an improving feasible direction. Let $I = \{i: g_i(x) = 0\}$ after a nonzero $\hat{P}$ has been found.

Step 2: If $X = \bigcap_{i \in I} [g_i(x) \cap \{R^{n-1} \text{ hyperplane}\}]$ is already linear, use the matrix $\hat{P}$ in the following calculations. Otherwise, find a coordinate change such that X becomes

$$Y = \{y: a^t y + c = 0\}$$

Find $\nabla f_{\text{new}}[y(x)] \mid y(x_u)$ and convert the problem into y coordinates. Form the projection matrix $P = I - a^t(aa^t)^{-1}a$ and go to *Step 3* after forming $d_u = P\nabla f_{\text{new}}[y(x)] \mid y(x_u)$.

Step 3: Let h_u be a solution to

$$\min f(z_u + hd_u) \text{ where } \begin{aligned} z_u &= x_u \quad \text{if in } x \text{ coordinates and} \\ z_u &= y_u \quad \text{if in } y \text{ coordinates} \\ 0 &\le h \le h_{\max} \end{aligned}$$

where $h_{\max}$ is determined so that the problem remains feasible. Let $z_{u+1} = z_u + hd_u$, convert into x coordinates, if necessary, and return to *Step 1*.

For more complex problems involving more than one binding constraint, the coordinate changes must be automated and checks made on the neighborhood of validity of the transformations. Application to other primal methods can also be made using the same types of arguments.

3.3. Sensitivity analysis and unfoldings

In Section 2, we noted that a universal unfolding of a smooth function $f(x)$ represents the most general type of smooth perturbation to which f can be subjected and that the number of terms needed to characterize all such perturbations equals codim f. Furthermore, if $u_1(x), \ldots, u_c(x)$ represent a basis for the Jacobian ideal $m_n / \nabla(f)$, then the $\{u_i\}$ also represent a basis for the space of all such perturbations. Since perturbations in the objective function and/or constraints lie at the heart of sensitivity analysis for nonlinear optimization, it seems reasonable to conjecture that the concepts of unfolding and transversality should be of use in characterizing various issues arising in the sensitivity analysis of nonlinear programs. Here we shall indicate two different directions to be pursued:

(1) Constraint qualification conditions.
(2) Objective function stabilization and examination of the stability of the dual algorithms discussed above.

3.4. Transversality and the Kuhn–Tucker conditions

As an indication of how singularity theory arguments can be employed to study constraint perturbations, let us examine the classical Kuhn–Tucker conditions using transversality arguments.

The Kuhn–Tucker necessary conditions play an important role in the theoretical development of mathematical programming. These conditions were derived from a more general set of conditions, called the Fritz John conditions, by assuming that a constraint qualification is in effect. Both the Fritz John and Kuhn–Tucker conditions are necessary for x^* to be an optimal solution of the constrained optimization problem. One of the most widely used constraint qualifications is that the gradients of the binding constraints at the point x^* be linearly independent.

In singularity theory, the concept of a transverse intersection between two manifolds is a cornerstone for structural stability arguments. One definition of a transverse intersection at a point is that no vector is perpendicular to the tangent spaces of both manifolds simultaneously (Poston and Stewart, 1978). Since the gradient vector of a manifold at a point will also be the normal vector to the tangent hyperplane at that point, it follows that the gradient vectors of two intersecting manifolds will both be collinear if and only if the intersection is transverse. Furthermore, and more importantly, the Thom Isotropy Theorem (Poston and Stewart, 1978) states that transverse crossings are structurally stable. This means that small perturbations of the constraints around a Kuhn–Tucker point would not change the geometry of the intersection much. In fact, the original constraint configuration can be recovered by a smooth coordinate change around the point of interest.

Let us consider an example demonstrating the structural instability of a nontransverse crossing. In the example, the following definition of a transverse crossing is used.

Definition 1.

Two manifolds, R and S, embedded in R^n intersect *transversally* if

(1) $R \cap S = \phi$

or

(2) $\operatorname{codim}(T_x R) + \operatorname{codim}(T_x S) < n$ and $\operatorname{codim}(T_x R) + \operatorname{codim}(T_x S) = \operatorname{codim}(T_x R \cap T_x S)$

where T_x is the tangent space at x.

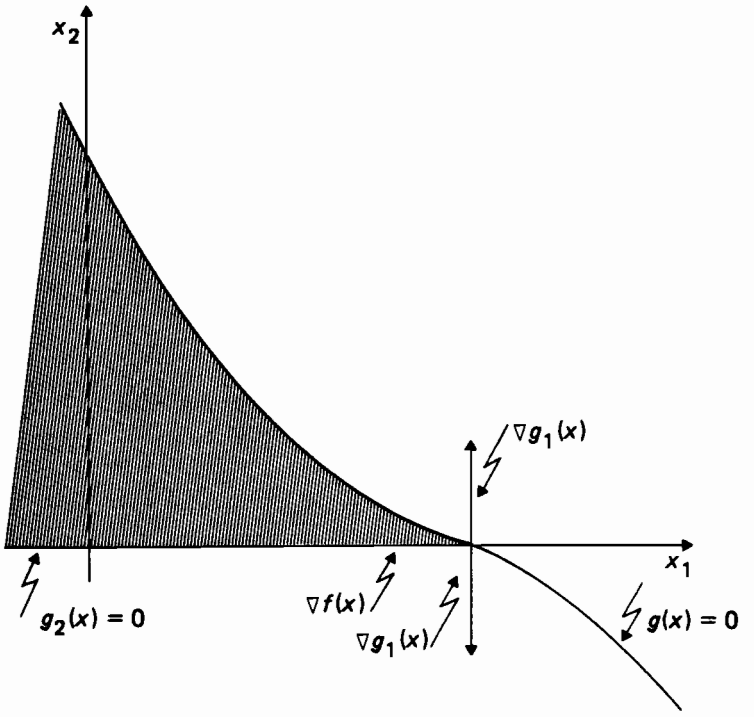


Figure 4 Example of a nontransverse constraint crossing (from Figure 4.5, Bazaraa and Shetty, 1979).

Example (from Bazaraa and Shetty, 1979; see *Figure 4* for geometry)

$$\min \; f(x_1,x_2) = -x_1$$

subject to

$$x_2 - (1-x_1)^3 \leq 0$$

$$-x_2 \leq 0$$

$\nabla f = (-1,0)$	at (1,0):	$\nabla f\,(1,0) = (-1,0)$	
$\nabla g_1 = [-3(1-x_1^2),1]$		$\nabla g_1\,(1,0) = (0,1)$	binding
$\nabla g_2 = (0,-1)$		$\nabla g_2\,(1,0) = (0,-1)$	binding

The gradients of the binding constraints are not linearly independent. Checking the Kuhn–Tucker conditions:

$$\nabla f + u_1 \nabla g_1 + u_2 \nabla g_2 = 0$$

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} + u_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 0$$

$$u_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} 0 = 1 \\ u_1 = u_2 \end{matrix}, \text{ inconsistency,}$$

showing that the Kuhn–Tucker conditions do not hold.

3.5. Transversality

Both g_1 and g_2 will be embedded in R^3 so

$$T_x\, g_1(x) = \{(x_1,x_2,x_3): x_2 - 3(1-x_1)^2 x_1 = x_2 - 3(1-x_1)^2\, x_1\}$$

$$T_x\, g_2(x) = \{(x_1,x_2,x_3): x_2 = x_2\}$$

At (1,0):

$$T_x\, g_1(x) = \{x_1,x_2,x_3): \; x_2 = 0\}, \text{ thus } T_x\, g_1(x) \cap T_x g_2(x) = T_x\, g_1(x)$$

$$T_x\, g_2(x) = \{(x_1,x_2,x_3): x_2 = 0\}$$

$$\text{codim } T_x\, g_1(x) = 1$$

$$\text{codim } T_x\, g_2(x) = 1$$

$$\text{codim } [T_x\, g_1(x) \cap T_x\, g_2(x)] = 1$$

Thus,

$$\text{codim } T_x\, g_1(x) + \text{codim } T_x\, g_2(x) \neq \text{codim } [T_x\, g_1(x) \cap T_x\, g_2(x)]$$

so the intersection is nontransverse.

If the cubic constraint is perturbed slightly,

$$g_1(y_1,y_2) = y_2 - (1-y_1)^3 + \varepsilon$$

then $T_x g_1(x) \cap T_x g_2(x)$ at (1,0) will be the empty set, so the intersection is, by definition, transverse. At their point of intersection, $x = (1+\varepsilon,0)$, so

$$T_x\, g_1(x) = \{(x_1,x_2,x_3): \; x_2 + 3\varepsilon^2\, x_1 = 3\varepsilon^2\}$$

$$T_x\, g_2(x) = \{(x_1,x_2,x_3): x_2 = 0\}$$

and $T_x\, g_1(x) \cap T_x\, g_2(x) = \{(x_1,x_2,x_3):\ x_1 = 1\}$ will be a line in R^3. Thus,

$$\text{codim}[T_x\, g(x)] = 1$$

$$\text{codim}[T_x\, g(x)] = 1$$

$$\text{codim}[T_x\, g_1(x) \cap T_x\, g_2(x)] = 2$$

so $\text{codim}[T_x\, g_1(x)] + \text{codim}[T_x\, g_2(x)] = \text{codim}[T_x\, g_1(x) \cap T_x\, g_1(x)]$ and the intersection is transverse.

The unfolding concept can also be of use in sensitivity analyses in the following manner. As an unparametrized function, the objective function $f(x)$ may be unstable with regard to small perturbations (i.e., the qualitative character of the critical points of f may change as a result of small changes in f). This is clearly a bad situation as far as the credibility of the results obtained from such an optimization is concerned. However, if codim $f = c$, an unfolding of f involving at least c parameters will be stable relative to all structural perturbations in the sense that if $f(x) + p(x)$ is a perturbation of f, then the behavior of $f(x) + p(x)$ near its critical points can be captured by varying the parameters in a universal unfolding of f. As already noted, the elements $u_1(x), \ldots, u_c(x)$, forming a basis for $m_n / \nabla(f)$, constitute a basis for exactly the type of perturbations we need to stabilize f.

Unfolding can also be of use for studying the stabililty of the dual optimization algorithms, which require the formation of a surrogate objective function using a computational parameter. For example, the augmented Lagrangian method mentioned above requires the use of a parameter ρ. These parameterized functions can be studied to learn what types of objective functions and constraints may lead to surrogate objective functions that are structurally unstable, and which may behave badly as the computational parameter is varied.

These ideas can be illustrated by considering the standard linear programming problem. In a sense, a linear objective function is the linearization of a general nonlinear $f(x)$, since no real-world process even generates a completely linear potential.

Definition 2

f is *structurally stable* if, for sufficiently small smooth perturbation functions p, the critical points of f and $f + p$ remain within the same neighborhood and have the same type (max, min, saddle, etc.).

Consider the linear program:

$$\max f(x) = c^t x$$

subject to

$$Ax \leq 0$$
$$x \geq 0$$

Note that the Hessian matrix of f will be identically zero for all x, so that a linear program has a maximum only by virtue of the constraints.

If a small linear perturbation is added to the objective function:

$$\max f(x) = (c^t + \varepsilon^t)x \qquad \varepsilon_i \ll 1 \qquad i = 1,2,\ldots,n$$

subject to

$$Ax \leq 0$$
$$x \geq 0$$

the isoclines of the objective function on the x hyperplane might shift so that the set of isoclines leaves the feasible region at a completely different extreme point of the convex hull of constraints. Thus, the linear programming problem is not even stable with respect to linear perturbations.

In contrast, it is known that Morse (i.e., quadratic) extrema are the only structurally stable types for nonparameterized functions, although for parameterized functions the situation is different. Similarly, since adding a small perturbation to a Morse function does not drastically change the location of the unconstrained extremum, the location of the constrained extremum also should not change too much, since the constrained extremum usually occurs where the constraints are tangent to the isoclines of $f(x)$.

As a final note, the computational implications of the above discussion are not by any means as dire as might seem. While the general nonlinear programming problem is computationally difficult, numerical methods for quadratic programs, both constrained and unconstrained, are well developed. In fact, since Morse functions are the only structurally stable types of smooth unparametrized functions, a case could be made for transforming even nonquadratic nonlinear programs into quadratic form using the diffeomorphic coordinate changes guaranteed by singularity theory. Thus, a quadratic program represents, in a certain sense, the canonical problem for mathematical programming.

Acknowledgment

Much of the work represented here arose during the course of numerous conversations with J. Kempf who, in particular, is responsible for the example of the use of the gradient projection method.

References

Arnol'd, V.I. (1981), *Singularity Theory* (Cambridge University Press, Cambridge, UK).

Bazaraa, M.S. and Shetty, C.M. (1979), *Nonlinear Programming* (John Wiley and Sons, New York).

Casti, J. (1984), *System Similarities and Natural Laws*, Working Paper WP-84-1 (International Institute for Applied Systems Analysis, Laxenburg, Austria).

Gibson, G. (1979), *Singular Points of Smooth Mappings* (Pitman, London).

Lu, Y.C. (1976), *Singularity Theory and an Introduction to Catastrophe Theory* (Springer, New York).

Milnor, J. (1973), *Morse Theory* (Princeton University Press, Princeton, NJ).

Poston, T. and Stewart, I. (1978), *Catastrophe Theory and Its Applications* (Pitman, London).

Modeling, Approximation, and Complexity of Linear Systems

J. C. Willems
Mathematics Institute, University of Groningen, The Netherlands

1. Introduction

The purpose of this paper is to give an extended summary of the material developed by Willems (forthcoming, a,b,c). In this sequence of papers we have outlined a theory that studies, in a broad context, the connection between various representations of dynamical systems and procedures for obtaining mathematical models on the basis of observed data.

2. Mathematical Models

In studying a phenomenon in the language of mathematics we first need to identify its features – its ***attributes*** – as elements of a set $\mathcal{S}$. A ***mathematical model*** is a law that excludes the occurrence of certain attributes. Hence, a mathematical model is a subset $\mathcal{M}$ of $\mathcal{S}$. Most phenomena will be described by attributes that consist of a product set

$$\mathcal{S} = \prod_{a \in A} \mathcal{S}_a$$

Interesting laws express the fact that variables are related in some way (as force/position, price/demand, electric/magnetic fields, etc.).

3. Dynamical Systems

When studying dynamical systems we have attributes that change as a function of time. This yields the following formal definition. A ***dynamical system*** Σ is defined by a triple $\Sigma = \{\mathbb{T}, \mathbb{W}, \mathcal{B}\}$ with $\mathbb{T} \subset \mathbb{R}$ the ***time axis***, $\mathbb{W}$ the ***signal alphabet***, and $\mathcal{B} \subset \mathbb{W}^{\mathbb{T}}$ the ***behavior***. Thus, a dynamical system consists of a family $\mathcal{B}$ of maps $w(.): \mathbb{T} \to \mathbb{W}$. If $w \in \mathcal{B}$, we think of w as being consistent with the laws governing the system. In this paper we only consider dynamical systems with $\mathbb{T} = \mathbb{Z}$.

Important properties of dynamical systems $\Sigma = \{\mathbb{Z}, \mathbb{W}, \mathcal{B}\}$ are:

(1) *Linearity:* if W is a vector space and B is a linear subspace of W^T
(2) *Time-invariance:* if $\sigma B = B$ (σ denotes the left shift).
(3) *Completeness:* if $\{w \in B\} \Longleftrightarrow \{w|_{[t_0,t_1]} \in B|_{[t_0,t_1]}$ for all $t_0,t_1 \in T\}$.
(4) *Memoryless:* if B is closed under concatenation.
(5) *Autonomous:* if for all t there exists a partial map:

$$f_t: W^{(-\infty,t)} \rightarrow W^{[t,\infty)}$$

such that

$$\{w \in B\} \Longleftrightarrow \{w|_{[t,\infty)} = f(w)|_{(-\infty,t)}\}$$

In the sequel we restrict attention to time-invariant systems, using the following notation:

$$B^-: = \{w^-: (-\infty,0) \rightarrow \mathbf{R}^q \mid \exists\, w^+ \quad w^- \cdot w^+ \in B\}$$
$$B^+: = \{w^+: [0,\infty) \rightarrow \mathbf{R}^q \mid \exists\, w^- \quad w^- \cdot w^+ \in B\}$$

Here $\cdot$ denotes concatenation. In terms of this notation Σ is memoryless if and only if $B = B^- \cdot B^+$ and Σ is autonomous if and only if B is the graph of map $f: B^- \rightarrow B^+$.

4. AR Systems

The following class of systems play an important role in system theory. Let $R \in \mathbf{R}^{g \times q}[s]$[: = the $(g \times q)$ real matrix polynomials] and consider $\Sigma(R)$: = $\{Z, \mathbf{R}^q, B(R)\}$ with $B(R)$: = $\{w: Z \rightarrow \mathbf{R}^q \mid R(\sigma)w = 0\}$. We will call such systems *AR systems*. Obviously $\sigma(R)$ is linear, time-invariant, and complete.

Proposition 1

The following are equivalent:

(1) $\Sigma = \{Z, \mathbf{R}^q, B\}$ is linear, time-invariant, and complete.
(2) B is a linear, shift invariant, and closed subspace of $(\mathbf{R}^q)^Z$ (equipped with the topology of pointwise convergence).
(3) There exists a polynomial matrix R such that $\Sigma = \Sigma(R)$, i.e., Σ is an AR system.

An important special class of AR systems are the *reachable* systems defined by the property $B^+(R) = B_c^+(R)$: = $\{w^+ \in B^+ \mid \exists\, w^- \in B^-$ such that $w^- \cdot w^+ \in B(R)$ and $w^-(t) = 0$ for t sufficiently small$\}$. It can be shown that $\{B(R)$ is reachable$\} \Longleftrightarrow \{\dim \ker R(s) =$ constant for $s \in \mathbb{C}\}\{B(R) = (B \cap$ compact$)^{\text{closure}}\}$. Here *compact* denotes the maps $w: Z \rightarrow \mathbf{R}^q$ with compact support, and the *closure* is to be understood in the topology of pointwise

convergence.

The other extreme are the *autonomous* systems. Thus $\{\mathcal{S}(R)$ is autonomous$\} \iff \{B_c^+(R) = 0\}$ $\{\dim \ker R(s) = \{0\}$ for almost all $s \in \mathbb{C}\}$.

A number of further special structures of AR systems are described next.

4.1. Input/output (I/O) systems

These are specified by two polynomial matrices P, Q with $\det P \neq 0$ and $P^{-1}Q$ a proper rational function which specifies B by means of the equations

$$P(\sigma)\mathbf{y} = Q(\sigma)\mathbf{u},\ \mathbf{w} = \mathrm{col}(\mathbf{u},\mathbf{y})$$

We denote these systems as $\Sigma(P;Q)$.

4.2. State space systems

A state space system in this category is specified by four matrices ($A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{p\times n}$, $D \in \mathbf{R}^{p\times m}$) and defines the behavior by means of the equations:

$$\sigma\mathbf{x} = A\mathbf{x} + B\mathbf{u} \qquad \mathbf{w} = C\mathbf{x} + D\mathbf{u}$$

Its external behavior is $B(A,B,C,D) = \{\mathbf{w} \mid \exists\ \mathbf{x},\mathbf{u}$ such that $\sigma\mathbf{x} = A\mathbf{x} + B\mathbf{u}$, $\mathbf{w} = C\mathbf{x} + D\mathbf{u}\}$. If the driving input is part of $\mathbf{w}$:

$$\sigma\mathbf{x} = A\mathbf{x} + B\mathbf{u} \qquad \mathbf{y} = C\mathbf{x} + D\mathbf{u} \qquad \mathbf{w} = \mathrm{col}(\mathbf{u},\mathbf{y})$$

then we denote this system as $B_{i/s/o}(A,B,C,D)$.

Proposition 2

It is possible to prove that the following are equivalent:

(1) $B = B(R)$ for some polynomial matrix R.
(2) $\exists\ P,Q$ with $P^{-1}Q$ proper and a permutation matrix π such that $B = \pi B(P;Q)$.
(3) $\exists$ matrices A,B,C,D such that $B = B(A,B,C,D)$.
(4) $\exists$ matrices A,B,C,D and a permutation matrix π such that $B = \pi B_{i/s/o}(A,B,C,D)$.

A state space system (A,B,C,D) will be said to be *minimal* if among all such systems with a given external behavior, n and m are minimal (minimality

of n and m are compatible). It can be shown that (A,B,C,D) is minimal if and only if

(1) $\ker D = \{0\}$.
(2) $A\mathbf{R}^n + \text{im } B = \mathbf{R}^n$.
(3) $[A - BD^*C, C \pmod{\text{im } D}]$ is observable.

Here D^* denotes the left inverse of D.

We can associate reachability and autonomy to its state space version. Indeed, {$\mathfrak{B}(R)$ is reachable} $\Longleftrightarrow$ {it allows a state space representation with (A,B) reachable} $\Longleftrightarrow$ {all its minimal representations have reachable (A,B)s}. {$\mathfrak{B}(R)$ is autonomous} $\Longleftrightarrow$ {it allows a state space representation with $m = 0$} $\Longleftrightarrow$ {all its minimal representations have $m = 0$}.

5. Recapitulation

As we have seen there are three basic "numerical" descriptions of the behavior of linear time-invariant complete systems:

(1) *AR models:* $R(\sigma)\mathbf{w} = 0$.
(2) *I/O models:* $P(\sigma)\mathbf{y} = Q(\sigma)\mathbf{u}$ $\quad w = (\mathbf{u},\mathbf{y})$.
(3) *State models:* $\sigma\mathbf{x} = A\mathbf{x} + B\mathbf{u}$ $\quad \mathbf{w} = C\mathbf{x} + D\mathbf{u}$.

There is also the combination of the latter two:

(4) i/s/o models: $\sigma\mathbf{x} = A\mathbf{x} + B\mathbf{u}$ $\quad \mathbf{y} = C\mathbf{x} + D\mathbf{u}$ $\quad \mathbf{w} = \text{col}(\mathbf{u},\mathbf{y})$.

This richness of meaningful versions of the same object yields immediately a number of questions:

(1) Questions of *representation*.
(2) Questions of *equivalence*.
(3) Questions of *canonical forms*.

Some of these problems are classical system theoretic questions, some of them appear here in a somewhat new light. They are discussed in some detail in the references.

6. Most Powerful Models

Let, in the language of Section 1, S be the set of possible *attributes* of a phenomenon. A *model* is a subset M of S. A *model set* $\mathfrak{M}$ is a subset $\mathfrak{M}$ of 2^S,

the set of all subsets of S. We will call a model M_1 *more powerful* than M_2 if $M_1 \subset M_2$; the more a model forbids, the better it is. Assume now that we have made a number of measurements of the phenomenon. This yields a set of attributes which, evidently, are possible. Hence a *measurement* is a subset Z of M. We will call a model M *unfalsified* by the measurement Z if $Z \subset M$. Let the model set $\mathcal{M}$ and the measurement Z be given. We will call M^* the *most powerful unfalsified model* if $Z \subset M^* \in \mathcal{M}$ and if $\{Z \subset M \in \mathcal{M}\} \Longrightarrow \{M^* \subset M\}$. Of course, M^* need not exist. It will if $S \in \mathcal{M}$ and if $\mathcal{M}$ has what we call the *intersection property*, meaning that an arbitrary intersection of subsets of $\mathcal{M}$ is again in $\mathcal{M}$.

In modeling q-dimensional time series we have $S = (\mathbb{R}^q)^{\mathbb{Z}}$. Hence, a model is a behavior $B \subset (\mathbb{R}^q)^{\mathbb{Z}}$, a dynamical system $\{\mathbb{Z}, \mathbb{R}^q, B\}$. Now if the model set $\mathcal{M}$ consists of all linear time-invariant complete systems, then $\mathcal{M}$ will obviously have the intersection property. Indeed, following *Proposition 1*, $\mathcal{M}$ then consists of all linear, shift invariant, closed subspaces of $(\mathbb{R}^q)^{\mathbb{Z}}$, which clearly yields the intersection property. From there it follows that if any family of time series is given, there will exist a most powerful unfalsified linear time-invariant finite dimensional system explaining Z.

Consider now the following concrete problem:

> Find, for an observed time series $\tilde{\mathbf{w}}: Z \longrightarrow \mathbb{R}^q$ the most powerful linear time-invariant complete dynamical system unfalsified by $\tilde{\mathbf{w}}$.

Denote this behavior by $B^*(\tilde{\mathbf{w}})$. Now, we may want to obtain $B^*(\tilde{\mathbf{w}})$ either in AR, in I/O, or in state-space form. In Willems (forthcoming b) several algorithms have been outlined to compute:

$$\tilde{\mathbf{w}} \longrightarrow B^*(\tilde{\mathbf{w}}) \quad \begin{matrix} R^*_{\tilde{\mathbf{w}}} \\ [A^*(\tilde{\mathbf{w}}), B^*(\tilde{\mathbf{w}}), C^*(\tilde{\mathbf{w}}), D^*(\tilde{\mathbf{w}})] \end{matrix}$$

Of course, if we consider real data $\tilde{\mathbf{w}}$ obtained with uncertainty as round-off error, etc., then we will almost surely obtain the trivial model $B^*(\tilde{\mathbf{w}}) = (\mathbb{R}^q)^{\mathbb{Z}}$. Hence, in trying to avoid falsification we will end up with a model that explains everything and hence teaches us nothing about the underlying phenomenon. Consequently, we are forced either to accept falsified models or to change the model set so that it can cope with uncertainty. This latter solution is what is advocated by statistics: in assuming that the mechanism which produces the data is driven in part by a stochastic process, we end up with a situation where, in principle, every data set is possible and where falsification becomes a moot point. Certainly, there are many instances where this is an eminently reasonable approach. However, it seems to us that in such areas as (adaptive) control, econometrics, signal processing, etc., the philosophy of statistics is much overused since it serves basically as a panacea for treating uncertainty without questioning whether or

not the basic premises underlying statistics (random sampling) are satisfied or not. The crucial problem in modeling on the basis of raw data is most often not the fact that the data is noisy, but that the modeler has consciously decided to explain the data by means of a model in a model set that is incapable of capturing the complexity of the underlying (nonlinear, time-varying, high dimensional, etc.) reality.

7. Complexity and Misfit

Instead of using most powerful in a set-theoretic sense, introduce complexity as a quantitative feature of the power of a model. Thus, the *complexity* will be a mapping from the model set into a partially ordered space, the complexity space $c: \mathcal{M} \to \mathbf{C}$. Instead of using falsification as an absolute notion, introduce misfit as a degree of falsification. Thus, the *misfit* will be a mapping from the model set and the measurement into a partially ordered space, the error level space:

$$\varepsilon: \mathcal{Z} \times \mathcal{M} \to \mathcal{E}$$

Here $\mathcal{Z} \subset 2^{\mathcal{S}}$ is a set of possible measurements. It is logical to assume that $\{\varepsilon(Z,M) = 0\} \Longleftrightarrow \{Z \subset M\}$.

Thus, if $Z \in \mathcal{Z}$ is observed and $M \in \mathcal{M}$ is postulated as a model for explaining Z, the modeler will have to weigh $c(M)$ against $\varepsilon(Z,M)$, since keeping both small is a conflicting objective.

8. The Complexity for Linear Systems

For the linear time-invariant complete systems discussed in Sections 4 and 5 the following complexity function is proposed. Let $\mathbf{L} := \{\text{all linear, shift-invariant, closed subspaces of } (\mathbb{R}^q)^{\mathbb{Z}}\}$. For $B \in \mathbf{L}$ define $B_t := B \mid [0,t]$. Let $C = [0,1]^{\mathbb{Z}_+}$ equipped with the partial ordering of pointwise domination. Define the complexity

$$c: \mathbf{L} \to [0,1]^{\mathbb{Z}_+} \text{ by } c_t(B) := \frac{\dim B_t}{q(t+1)}$$

This complexity measure can be nicely expressed in terms of AR, I/O, or i/s/o specifications of B. We concentrate on the latter. It can be shown that there is a one-to-one relation between $c(B)$ and $(m, \nu_1, \nu_2, \ldots, \nu_q)$, with m the number of inputs and $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_q, \sum_{i=1}^{q} \nu_i = n$, the *observability indices* of any minimal i/s/o representation of B.

9. The Misfit for Linear Systems

For the linear time-invariant complete systems discussed in Sections 4 and 5 the following misfit function is proposed. Let $K \subset (\mathbb{R}^q)^{\mathbb{Z}}$ be an inner product space and let $Z = \{\tilde{w} : Z \to \mathbb{R}^q \mid \text{each component } \tilde{w}_i \in K\}$. Take $E = (\mathbb{R}_+)^{\mathbb{Z}_+}$ and define the misfit ε: $Z \times \mathbf{L} \to (\mathbb{R}_+)^{\mathbb{Z}_+}$ by

$$\varepsilon_t(\tilde{w},B) = \max_{a \in N'_t} \frac{\|a(\sigma)\tilde{w}\|_K}{\|a\|_{L_1^*}}$$

where $N'_t \subset L_t^*$: $= \{a \subset \mathbb{R}^{1 \times q}[s] \mid \partial(a) \leq t\}$ ($\partial :=$ degree) is defined as follows:

$$N'_0 = B_0^\perp \qquad N'_t = B_t^\perp \cap (B_{t-1}^\perp + \sigma^* B_{t-1}^\perp)^\perp$$

Here σ^* denotes the right shift while $B_t^\perp$: $= \{a \in L_t^* \mid a^*(\sigma)B = 0\}$. Thus, ε_t is a measure for in how far the measured time series $\tilde{w}$ fails to corroborate the tth order lag AR relations postulated by accepting the model B.

10. Optimal Approximate Modeling

We pursue two methodologies for approximate modeling. Assume that S, M, Z, $c: \mathcal{M} \to C$, and $\varepsilon: Z \times \mathcal{M} \to E$ are given.

(1) Fix $c^{\text{adm}} \in C$, the *maximal admissible complexity*. Now compute the *optimal approximate model* $M^* \in \mathcal{M}$ such that it satisfies

- (a) $c(M^*) \leq c^{\text{adm}}$.
- (b) $\{M \in \mathcal{M}, c(M) \leq c^{\text{adm}}\} \Longrightarrow \{\varepsilon(Z,M^*) \leq \varepsilon(Z,M)\}$.
- (c) $\{M \in \mathcal{M}, c(M) \leq c^{\text{adm}}, \varepsilon(Z,M) = \varepsilon(Z,M^*)\} \Longrightarrow \{c(M^*) \leq c(M)\}$.

(2) Fix $\varepsilon^{\text{tol}} \in E$, the *maximal tolerated misfit*. Now compute the *optimal approximate model* $M^* \in \mathcal{M}$ such that it satisfies

- (i) $\varepsilon(Z,M^*) \leq \varepsilon^{\text{tol}}$.
- (ii) $\{M \in \mathcal{M}, \varepsilon(Z,M) \leq \varepsilon^{\text{tol}}\} \Longrightarrow \{c(M^*) \leq c(M)\}$.
- (iii) $\{M \in \mathcal{M}, \varepsilon(Z,M) \leq \varepsilon^{\text{tol}}, c(M^*) = c(M)\} \Longrightarrow \{\varepsilon(Z,M^*) \leq \varepsilon(Z,M)\}$.

Using slight variations of the complexity measure discussed in Section 8 and the misfit discussed in Section 9 we have developed in Willems (forthcoming c) algorithms and examples which give M^* following the above methodologies.

In the philosophy of these two methodologies it is quite acceptable to produce models that are falsified by the data provided, so long as the complexity and/or misfit are not excessively high. Of course, this philosophy

may be useful in a stochastic context as well: coding information with a limited distortion rate is a reasonable basis for stochastic modeling. More yet, complexity considerations could in some cases serve as a conceptual justification of stochastics in the sense that in circumstances where stochastic inputs and/or sampling are hard to justify from the experimental set-up, stochastic models may be the least complex ones that corroborate the data within a given misfit level. Think, for example, about the rolling of an honest dice. A deterministic model, which beats the misfit of the simple equal distribution model postulated on the basis of the principle of insufficient reason, will undoubtedly have to be extremely complex.

References

Willems, J.C. (forthcoming a), From time series to linear system. Part I: Finite dimensional linear systems, submitted for publication.

Willems, J.C. (forthcoming b), From time series to linear system. Part II: Exact modeling, submitted for publication.

Willems, J.C. (forthcoming c), From time series to linear system. Part III: Approximate modeling, in preparation

V. BIOLOGICAL AND SOCIAL APPLICATIONS

Cycling in Simple Genetic Systems: II. The Symmetric Cases

E. Akin
Department of Mathematics, City College, 137th Street and Convent Avenue, NY 10031, USA

Dedicated to the memory of Charles Conley

1. Introduction

Sewall Wright's adaptive landscape is the picture we all use to visualize the dynamics of evolution, at least at the microlevel. Imagine a flat plane each point of which represents a genetic state of the gene pool of a population. Upon this plane is erected a continuous topography whose height above a point describes the degree of adaptedness, or fitness, associated with the corresponding genetic state. The dynamic assumption is that natural selection moves the population upward, in the direction of increasing fitness, with equilibria at local maxima or more general critical points of the fitness function.

This picture is wonderfully versatile in allowing us to imagine the interaction between selection and other evolutionary effects. We model slow environmental effects by geological changes of the topography. We can thus contrast the slow tracking by a population of the gradual movement of a peak with the relatively rapid reorganization subsequent to the erosion of a peak of intermediate height into the slope of a still higher peak. Mathematical analysis of the corresponding gradient bifurcation merely formalizes this picture. We model genetic drift, the sampling effect due to finite population size, by superimposing a Brownian motion. We thus see both the cost of drift as the local optimum equilibrium point of pure selection is expanded to a distributed blot of nearby points and also the benefit in freeing a population from the suboptimal destiny of a local maximum of intermediate height. The great value to biology of the analytical elaboration of this model by Kimura and his associates comes not from enhanced imaginative understanding but instead from his subtle deployment of rate estimates in the neutral gene controversy.

The mathematical inspiration for the adaptive landscape is R.A. Fisher's *Fundamental Theorem of Natural Selection*. Consider a large Mendelian population with a finite list I of gametic genotypes. The state of the population is the distribution vector $\mathbf{p}$ of gametic genotypes. So the state space is the unit simplex Δ consisting of nonnegative vectors whose components sum to one. Selection is modeled by the system of differential equations:

$$\frac{dp_i}{dt} = p_i(m_i - \bar{m}) \qquad (i \in I) \tag{1}$$

Here $m_i = \Sigma_j\, p_j m_{ij}$ and $\bar{m} = \Sigma_i\, p_i m_i = \Sigma_{i,j}\, p_i p_j m_{ij}$ with m_{ij} the fitness (assumed constant) of the zygotic genotype ij. Because the zygote pair ij is unordered the fitness matrix is symmetric.

The function $\bar{m}$ of p is the mean fitness of the population and Fisher's Theorem says that $\bar{m}$ tends to increase. To see this check that:

$$\frac{1}{2}\frac{d\bar{m}}{dt} = \Sigma_i\, p_i\, (m_i - \bar{m})^2$$

The right side, the so-called additive variance, is nonnegative, vanishing only at equilibrium.

Extending Fisher's Theorem, Kimura's Maximum Principle says that the direction of the vector field of equation (1) is that of greatest increase of $\bar{m}$. Shahshahani (1979) has shown that the vector field is precisely the gradient of (0.5) $\bar{m}$ with respect to an appropriate Riemannian metric devised by Conley and himself. The metric is related to a measure of information due to R.A. Fisher and had earlier been introduced into biology in a somewhat different form by Antonelli and Strobeck (1977). An exposition of the geometry of this Shahshahani metric is given in Akin (1979).

There is a difference equation analogue of (1), which has also been widely studied. With one exception, which we note below, the results of the two kinds of models are completely parallel. However, the continuous time, vector-field version (1) is easier to analyze than the discrete time, mapping analogue.

The derivation of equation (1) assumes that an ij zygote produces upon maturity gametes of type i and j. In addition to neglecting mutation this means we consider I to be a set of alternative alleles at a single autosomal (i.e., not sex-linked) locus, or a system that behaves as such, e.g., alternate inversion types of a single autosomal chromosome. However, when several incompletely linked loci are admitted the equations must be corrected to allow for the possibility of recombination between paired chromosomes.

We are concerned here with the simplest case where recombination occurs: the two-locus–two-allele (TLTA) model. Following custom we label the two alleles of the first locus A and a, those of the second locus B and b, and number the four gamete types $1 = AB$, $2 = Ab$, $3 = aB$, $4 = ab$. When recombination is incorporated, equation (1) becomes:

$$\frac{dp_i}{dt} = p_i\, (m_i - \bar{m}) - Rd\, \varepsilon_i \quad (i = 1, \ldots, 4) \tag{2}$$

with $\varepsilon_1 = \varepsilon_4 = +1$ and $\varepsilon_2 = \varepsilon_3 = -1$. Here R is the (constant) recombination rate per unit time (i.e., crossovers per birth times the birth rate for double heterozygotes) and d is the difference measure of linkage defined by:

$$d = p_1 p_4 - p_2 p_3 \tag{3}$$

From the probability distribution p we compute the marginal distributions at the separate loci by:

$$p_A = p_1 + p_2 \qquad p_a = 1 - p_A = p_3 + p_4$$
$$p_B = p_1 + p_3 \qquad p_b = 1 - p_B = p_2 + p_4$$

That d measures linkage, i.e., dependence between loci, follows from the easily checked formulae:

$$p_1 = p_A p_B + d \qquad p_3 = p_a - d$$
$$p_2 = p_A p_b - d \qquad p_4 = p_a + d$$

The unit-sum constraint on the simplex Δ implies that the system is three dimensional. It is most natural to use one variable to describe the linkage, e.g., d. Notice that the additional recombination term in equation (2) vanishes precisely at the linkage equilibrium when $d = 0$.

Recombination generates new gametes only in the double heterozygote cases $ij = 14$ or 23, where all four alleles occur. For biological reasons it is usually assumed that these double heterozygotes have the same fitness (no "position effects"). The fitness matrix can then be displayed via a square table, *Table 1*.

Table 1

	aa	*Aa*	*AA*
BB	m_{33}	m_{13}	m_{11}
Bb	m_{34}	$m_{14} = m_{23}$	m_{12}
bb	m_{44}	m_{24}	m_{22}

Notice that adding a common constant to the entries of the fitness matrix has no effect on the equations. In the competition between gamete types it is a relative advantage that is important. So we can normalize the fitnesses by assuming that the central entry in *Table 1* is zero.

While equation (2) and its discrete time analogue have been widely applied to biological problems, the pure analysis of these systems has consisted of a series of shocks administered to the biological intuition nurtured upon the adaptive topography picture.

First, Moran (1964) showed that mean fitness need not always increase along solution paths of equation (2). To see this, begin with a selection matrix such that mean fitness has an isolated maximum at p^*, an interior distribution (i.e., $p_i^* > 0$ for $i = 1, \ldots, 4$) not in linkage equilibrium (i.e. $d \neq 0$ at p^*). Then p^* is a stable equilibrium for selection alone [equation (2) with $R = 0$], but is not an equilibrium when R is positive. Thus, for small positive

values of R, equation (2) will have a locally stable equilibrium displaced from the peak of $\bar{m}$ at p^*. Some approaches to this equilibrium will have to proceed downhill. Ewens (1969) showed that this kind of example requires epistasis defined as nonadditivity of fitness between loci.

Moran's result leaves open the possibility that equation (2) maximizes some "fitness function" other than $\bar{m}$. While selection alone maximizes mean fitness, recombination alone maximizes a function related to entropy (Akin, 1979, p 138; see also Kun and Lyubich, 1979). Perhaps some mixture of mean fitness and entropy will serve as a Ljapunov function or, following Ewens, perhaps some adjustment of $\bar{m}$ to account for epistasis will do. This hope was nourished by the belief that every solution path of equation (2) converges to equilibrium. Such convergent systems usually admit many Ljapunov functions, but which one is biologically meaningful?

Selection alone ($R = 0$) is the gradient of $0.5\bar{m}$ with equilibria at the critical points of the restriction of $\bar{m}$ to each open face of Δ; convergence is true in that case. Obvious when the critical points are isolated, the general – not obvious – result was proved by Lyubich *et al.* (1980) (see also Losert and Akin, 1983).

Once $R > 0$ even the pattern of equilibria becomes difficult to analyze. Attention focused upon the special cases of equation (2) invariant under the involution $\pi_0(p_1,p_2,p_3,p_4) = (p_4,p_3,p_2,p_1)$. This occurs when the selection matrix is origin symmetric when presented as in *Table 1*, i.e., $m_{33} = m_{22}$, $m_{13} = m_{24}$, etc. The fixed point set of π_0 is the segment

$$\mathrm{Fix}(\pi_0) = \{p \in \Delta: p_A = p_a = p_B = p_b = 0.5\}$$

For these symmetric cases $\mathrm{Fix}(\pi_0)$ is an invariant manifold and the restricted system is one dimensional and easy to analyze. Finally, Karlin and Feldman (1970) dqescribed all the equilibria, including the asymmetric ones away from the fixed point set. They used the discrete time analogue of equation (2) but the results are the same for (2) itself.

As far as I am aware the first observation that convergence to equilibrium is not inevitable occurs in Kun and Lyubich (1979). The authors' example is π_0 symmetric in discrete time and the barycenter $p^* = (0.25)(1, 1, 1, 1)$ [i.e., $p^* \in \mathrm{Fix}(\pi_0)$ and $d = 0$ at p^*] is an equilibrium for all values of R. It is locally stable for small $R > 0$. Restricting to the invariant segment one can compute that at about $R = 0.5$ the map becomes orientation reversing near p^* (the eigenvalue at the fixed point becomes negative). Stability is retained but the approach to equilibrium is oscillatory. Then, near $R = 0.75$ a period doubling bifurcation occurs (the eigenvalue moves below -1), p^* loses stability and is now flanked by a pair of points forming a stable orbit of period 2. For this periodic pair and its domain of attraction convergence to equilibrium does not occur. This "overshooting" behavior does not carry over from the difference equation model to (2) itself. However, nonconvergence can occur in the differential equations model as well.

Akin (1979) showed that Hopf bifurcation can occur in the family of selection-plus-recombination models. Thus, nontrivial periodic solutions are possible for equation (2). A Hopf bifurcation at an equilibrium happens when a parameter change causes a complex conjugate pair of eigenvalues to cross the imaginary axis. Because the selection vector field is a gradient its linearization is everywhere symmetric (with respect to the Shahshahani metric) and so has only real eigenvalues. Despite the entropy-like Ljapunov function the recombination field is not a gradient. The linearization is not symmetric except at linkage equilibrium and at points of $\mathrm{Fix}(\pi_0)$. Fix p^* at any other interior point. At p^* the linearization of the recombination field has a nonzero antisymmetric part. If position effects are allowed the family of selection matrices is rich enough to fix p^* as an equilibrium and to give any arbitrarily chosen symmetric part for the linearization of the combined field. It is then elementary to construct a one-parameter family of selection matrices so adapted to the antisymmetric map that the desired eigenvalue crossing occurs at p^*. In a more detailed study of the TLTA case (Akin, 1983, summarized in Akin, 1982), it was shown that position effects could be excluded and that limit cycles, locally stable periodic solutions, could appear. At about the same time Hastings (1981) showed that analogous Hopf bifurcations occur for the discrete time version of equation (2).

In order to digest this startling behavior the biologist requires a good stock of examples to chew upon. The proofs alone provide little nourishment for the imagination. In the next section I describe such a stock of examples. In using them one fixes the selection matrix and uses the recombination rate R as a parameter. Their symmetry allows us to restrict the three-dimensional system to a two-dimensional invariant plane where the orbit structure is easy to visualize. The key is to abandon π_0 in favor of a different involution.

As described in Akin (1983, p 42) the natural symmetry group for the class of TLTA models is isomorphic to the symmetry group of the square. In addition to the origin involution corresponding to π_0 there are two other conjugate pairs of involutions corresponding to the axis and diagonal symmetries. The involution corresponding to the $x = y$ diagonal is given by $\pi_+(p_1,p_2,p_3,p_4) = (p_1,p_3,p_2,p_4)$ with the two-dimensional fixed point set $\mathrm{Fix}(\pi_+) = \{p \in \Delta: p_2 = p_3\}$. We know in advance from results in Akin (1983) that π_+ invariant Hopf bifurcations occur and that the resulting cycles lie in $\mathrm{Fix}(\pi_+)$.

I must leave the elaboration of these examples for later publication or to the impatient reader who wishes to anticipate my rather ponderous engagement with computer graphics. Preliminary results suggest that the cycles occur for only a thin range of parameter values. After they appear they do not grow very large but are instead rather quickly eliminated by a saddle-crossing bifurcation.

2. π_+ Invariant TLTA Systems

Since we are to study systems on the unit simplex Δ in $\mathbf{R}^4$ invariant under $\pi_+(p_1,p_2,p_3,p_4) = (p_1,p_3,p_2,p_4)$, we begin by choosing coordinates related

to π_+. For convenience, we denote by p_+ the geometric mean of p_2 and p_3:

$$p_+ \equiv (p_2 p_3)^{0.5}$$

So $p_+ > 0$ when p_2 and p_3 are positive. This condition defines an open subset of Δ containing the interior distributions. When $p_+ > 0$ we define:

$$\begin{aligned} w &= \ln(p_2/p_3)^{0.5} = 0.5 \ln(p_2/p_3) = 0.5(\ln p_2 - \ln p_3) \\ u &= p_1/p_+ \qquad v = p_4/p_+ \end{aligned} \tag{4}$$

Thus, u and v are nonnegative while w can have any real value. To invert we note that:

$$\begin{aligned} e^w &= p_2/p_+ \qquad e^{-w} = p_3/p_+ \\ u &+ e^w + e^{-w} + v = 1/p_+ \end{aligned} \tag{5}$$

So the distribution vector $\mathbf{p}$ is (u, e^w, e^{-w}, v) divided by the sum of the components.

In terms of these coordinates our involutions become $\pi_0(w,u,v) = (-w,v,u)$ and $\pi_+(w,u,v) = (-w,u,v)$. Hence:

$$\begin{aligned} \mathrm{Fix}(\pi_0) &= \{\mathbf{p} : u = v \text{ and } w = 0\} \\ \mathrm{Fix}(\pi_+) &= \{\mathbf{p} : w = 0\}. \end{aligned} \tag{6}$$

The linkage disequilibrium variable is:

$$d = p_1 p_4 - p_2 p_3 = (uv - 1)p_+^2 \tag{7}$$

So $\{d = 0\}$ corresponds to the hyperbolic cylinder defined by $uv = 1$.

p_1 and p_4 are the natural two coordinates on $\mathrm{Fix}(\pi_+)$ and u and v are versions of these two. The third coordinate should measure deviation from the fixed point set, e.g., either $p_2 - p_3$ or $\ln p_2 - \ln p_3$, and w is a version of the latter.

By a π_+ invariant selection matrix we mean one whose *Table 1* representation is normalized by a central zero and is invariant under the $x = y$ diagonal symmetry. Such a matrix is a real 4×4 symmetric matrix satisfying equation (8).

$$\begin{aligned} m_{22} &= m_{33} \\ m_{12} &= m_{13} \\ m_{24} &= m_{34} \\ m_{14} &= m_{23} = 0 \end{aligned} \tag{8}$$

So we are left with five independent parameters: m_{11}, m_{22}, m_{44}, m_{12}, m_{24}. In the corresponding selection–recombination equations (2) the recombination rate R is the sixth parameter.

In rewriting equation (2) in our new coordinates we use the hyperbolic cosine function $\cosh w = 0.5\,(e^{w} + e^{-w})$ and the logarithmic average:

$$Q(a,b) \equiv \frac{a-b}{\ln a - \ln b} \qquad (a,b > 0) \tag{9}$$

Q is a smooth positive function that satisfies $Q(a,b) = Q(b,a)$ and $Q(a,a) = a$ (see, e.g., Akin, 1979, p 136).

On the open set $\{p_+ > 0\}$, equation (2) becomes

$$\begin{aligned} \frac{1}{p_+}\frac{dw}{dt} &= wQ(e^{w},e^{-w})\,[m_{22} + R(1-uv)] \\ \frac{1}{p_+}\frac{du}{dt} &= u[(m_{11}-m_{12})u + (2m_{12}-m_{22})\cosh w + (-m_{24})v] \\ &\quad + R(1-uv)(1 + u\cosh w) \\ \frac{1}{p_+}\frac{dv}{dt} &= v[(-m_{12})u + (2m_{24}-m_{22})\cosh w + (m_{44}-m_{24})v] \\ &\quad + R(1-uv)\,(1 + v\cosh w) \end{aligned} \tag{10}$$

Compute as follows:

$$\frac{dw}{dt} = 0.5\left[\frac{d\ln p_2}{dt} - \frac{d\ln p_3}{dt}\right] = 0.5\left[m_2 - m_3 - Rd\left[\frac{\varepsilon_2}{p_2} - \frac{\varepsilon_3}{p_3}\right]\right]$$

From equation (8):

$$m_2 - m_3 = m_{22}\,(p_2 - p_3) \quad (m_{21}p_1 = m_{31}p_1, \text{ etc.})$$

Also $\varepsilon_2 = \varepsilon_3 = -1$ and $d/p_2p_3 = d/p_+^2 = uv - 1$. Hence:

$$\frac{dw}{dt} = 0.5(p_2 - p_3)\,[m_{22} + R(1-uv)]$$

Then we rewrite $p_2 - p_3$ as:

$$2\,p_+[0.5(\ln(p_2/p_+) - \ln(p_3/p_+))]\left[\frac{p_2/p_+ - p_3/p_+}{\ln(p_2/p_+) - \ln(p_3/p_+)}\right]$$

The first bracketed factor is w and the second is $Q(e^w, e^{-w})$.

For du/dt (the argument for dv/dt is similar), notice:

$$\frac{du}{dt} = u\left[\frac{d\ln p_1}{dt} - 0.5\frac{d\ln p_2}{dt} - 0.5\frac{d\ln p_3}{dt}\right]$$

$$= u(m_1 - 0.5m_2 - 0.5m_3) - Rd\,\frac{p_1}{p_+}\left[\frac{\varepsilon_1}{p_1} - \frac{\varepsilon_2}{2p_2} - \frac{\varepsilon_3}{2p_3}\right]$$

From equation (8) again we compute that:

$$m_1 - 0.5m_2 - 0.5m_3 = (m_{11} - m_{12})p_1 + (2m_{12} - m_{22})\frac{p_2 + p_3}{2} + (-m_{24})p_4$$

Multiply and divide by p_+ to obtain the coefficient of u.

Finally, the recombination terms equal:

$$p_+ R\left[\frac{-d}{p_+^2}\right]\left[1 + \frac{p_1}{p_+}0.5\left[\frac{p_+}{p_2} + \frac{p_+}{p_3}\right]\right] = p_+R(1 - uv)(1 + u\cosh w) \quad \text{QED}$$

For convenience we rearrange the selection parameters, defining:

$$\begin{aligned} \alpha &= m_{11} - m_{12} \qquad \beta = -m_{24} \qquad \gamma = -m_{12} \\ \delta &= m_{44} - m_{24} \qquad \varepsilon = 2m_{12} + 2m_{24} - m_{22} \end{aligned} \tag{11}$$

which can be inverted via:

$$\begin{aligned} m_{12} &= -\gamma \qquad m_{24} = -\beta \qquad m_{11} = \alpha - \gamma \\ m_{44} &= \delta - \beta \qquad m_{22} = -\varepsilon - 2\beta - 2\gamma \end{aligned} \tag{12}$$

Proposition 1

For a π_+ invariant selection matrix, the system (2) is invariant under π_+. In particular, $\mathrm{Fix}(\pi_+) = \{p: p_2 = p_3\}$ is a two-dimensional invariant manifold. On the open subset of this two-cell where $p_2 = p_3 > 0$ we define

$u = p_1/p_2 = p_1/p_3$ and $v = p_4/p_2 = p_4/p_3$ so that $1/p_2 = 1/p_3 = u + 2 + v$. By introducing the time change:

$$\frac{d\tau}{dt} = p_2 = p_3 \tag{13}$$

Using the renaming equation (11) the restriction of the system to Fix (π_+) becomes:

$$\begin{aligned}\frac{du}{d\tau} &= (\varepsilon + 2\beta + \alpha u + \beta v)u + R(1 - uv)(1 + u) \\ &= [R + (\varepsilon + 2\beta + R)u + \alpha u^2] + [u(\beta - R - Ru)]v \equiv F(u,v) \qquad (14) \\ \frac{dv}{d\tau} &= (\varepsilon + 2\gamma + \gamma u + \delta v)v + R(1 - uv)(1 + v) \\ &= [R + (\varepsilon + 2\gamma + R)v + \delta v^2] + [v(\gamma - R - Rv)]u \equiv G(u,v)\end{aligned}$$

In particular, (u^*,v^*) is an equilibrium for equation (14) if and only if the corresponding point p^* is an equilibrium for equation (2). Two of the eigenvalues for p^* are those of (u^*,v^*) in equation (14) divided by the positive number $u^* + 2 + v^*$. The third eigenvalue is $\mu/(u^* + 2 + v^*)$ where

$$\mu \equiv R(1 - u^*v^*) - \varepsilon - 2\beta - 2\gamma \tag{15}$$

Proof

Clearly, equation (10) is invariant under $w \to -w$ and $w = 0$ is an invariant manifold. When $w = 0$, $\cosh w = Q(e^w, e^{-w}) = 1$. So equation (14) follows from (10) and the definitions (11) and (12). The eigenvalues before and after the time change are related by the constant $p_+ = 1/(u^* + 2 + v^*)$. By invariance, the third eigenvalue is just the coefficient of w in dw/dt, i.e., $R(1 - u^*v^*) + m_{22}$. QED

Notice that F is linear in v and G is linear in u. Thus, the locus of $F = 0$ is the graph of the implicitly defined function of u whose display is an easy exercise in curve sketching; similarly, for $G = 0$. The equilibria are the intersection points of the two graphs.

To generate examples we adapt the parameterization methods of Akin (1983). We work backward beginning with a point (u^*,v^*) and then specify values of the parameters $\alpha, \beta, \ldots,$ etc., so that (u^*,v^*) is a Hopf equilibrium, i.e., the eigenvalues are pure imaginaries. We have enough parameters that we can demand a specially nice form for the linearization. Notice that multiplying all of the parameters (including R) by a common positive constant changes the speed but not the solution paths. Thus, we obtain the ratios α/R, β/R, etc., as functions of the chosen point (u^*,v^*).

Proposition 2

Fix $u^*, v^* > 0$. With F and G defined by equation (14) there is a choice of values for $\alpha, \beta, \gamma, \delta, \varepsilon,$ and R unique up to a positive multiple, so that $F(u^*, v^*) = 0 = G(u^*, v^*)$ and at (u^*, v^*):

$$\begin{bmatrix} \frac{\partial F}{\partial u} & \frac{\partial F}{\partial v} \\ \frac{\partial G}{\partial u} & \frac{\partial G}{\partial v} \end{bmatrix} = \begin{bmatrix} 0 & \lambda \\ -\lambda & 0 \end{bmatrix} \tag{16}$$

The ratios α/R, β/R, γ/R, δ/R, ε/R, λ/R, and μ/R are rational functions of u^* and v^*. Furthermore, $\lambda = 0$ if and only if $u^* = v^*$ or $u^* v^* = 1$, while μ has the same sign as $1 - u^* v^*$.

Proof

For the computation we normalize by setting $R = 1$ and for the duration of the proof we omit the asterisks on u^* and v^*, so:

$$\begin{aligned} F &= [1 + (\varepsilon + 2\beta + 1)u + \alpha u^2] + [u(\beta - 1 - u)]v \\ \frac{\partial F}{\partial u} &= [(\varepsilon + 2\beta + 1) + 2\alpha u] + (\beta - 1 - 2u)v \\ G &= [1 + (\varepsilon + 2\gamma + 1)v + \delta v^2] + [v(\gamma - 1 - v)]u \\ \frac{\partial G}{\partial v} &= [(\varepsilon + 2\gamma + 1) + 2\delta v] + (\gamma - 1 - 2v)u \end{aligned}$$

The six equations we will use to determine the six parameters (including the new one, λ) are:

$$\frac{\partial F}{\partial v} = \lambda \tag{17}$$

$$\frac{\partial G}{\partial u} = -\lambda \tag{18}$$

$$F - u\frac{\partial F}{\partial u} = 0 \tag{19}$$

$$G - v\frac{\partial G}{\partial v} = 0 \tag{20}$$

$$\frac{\partial F}{\partial u} - \frac{\partial G}{\partial v} = 0 \tag{21}$$

$$\frac{\partial F}{\partial u} + \frac{\partial G}{\partial v} = 0 \tag{22}$$

Equations (17) and (18) determine β and γ:

$$\beta - 1 = u + \lambda u^{-1} \qquad \gamma - 1 = v - \lambda v^{-1} \tag{23}$$

Equation (19) says $1 - \alpha u^2 + u^2 v = 0$ and so we determine α from it and δ from equation (20):

$$\alpha = u^{-2} + v \qquad \delta = v^{-2} + u \tag{24}$$

Equation (21) gives:

$$2(\beta - \gamma) + 2(\alpha u - \delta v) + (\beta - 1)v - (\gamma - 1)u = 0$$

and can be rewritten using equations (23) and (24):

$$2(u - v + \lambda u^{-1} + \lambda v^{-1}) + 2(u^{-1} - v^{-1}) + \lambda u^{-1} v + \lambda v^{-1} u = 0$$

Multiply by uv and solve for λ:

$$\lambda = \frac{2(u - v)(1 - uv)}{2u + 2v + u^2 + v^2} \tag{25}$$

Equation (22) gives:

$$2\varepsilon + 2(\beta + \gamma) + 2 + 2(\alpha u + \delta v) + (\beta - 1 - 2u)v + (\gamma - 1 - 2v)u = 0$$

and can be rewritten using equations (23) and (24) as:

$$2\varepsilon + 2[u + v + 2 + \lambda(u^{-1} - v^{-1})] + 2 + 2(u^{-1} + v^{-1} + 2uv) + (\lambda u^{-1} - u)v + (-\lambda v^{-1} - v)u = 0$$

or

$$\varepsilon = 0.5\lambda u^{-1} v^{-1}(u - v)(2 + u + v) - (3 + u + v + u^{-1} + v^{-1} + uv)$$

Finally, we compute μ defined by equation (15), with $R = 1$:

$$\begin{aligned}\mu &= 1 - uv - \varepsilon - 2\beta - 2\gamma \\ &= 1 - uv - 0.5\lambda u^{-1} v^{-1}(u - v)(u + v + 2) + 3 + u + v \\ &\quad + u^{-1} + v^{-1} + uv + 2\lambda u^{-1} v^{-1}(u - v) - 4 - 2u - 2v \\ &= -0.5\lambda u^{-1} v^{-1}(u - v)(u + v - 2) + u^{-1} + v^{-1} - u - v\end{aligned}$$

Now apply equation (25) and the identities:

$$u^{-1} + v^{-1} - u - v = u^{-1}v^{-1}(1 - uv)(u + v)$$
$$2u + 2v + u^2 + v^2 = (u - v)^2 + 2u + 2v + 2uv$$

so that μ becomes:

$$\frac{1 - uv}{uv(2u + 2v + u^2 + v^2)} \{-(u - v)^2(u + v - 2) + (u + v)[(u - v)^2 + 2u + 2v + 2uv]\}$$

The expression in the braces is:

$$2(u - v)^2 + 2(u + v)^2 + 2uv(u + v) = 2[2u^2 + 2v^2 + uv(u + v)]$$

Thus, all of the factors in μ are positive except for $(1 - uv)$. QED

We summarize the results of our computations in *Table 2*.

Table 2

$$\lambda/R = 2(u^* - v^*)(1 - u^*v^*)/[2u^* + 2v^* + (u^*)^2 + (v^*)^2]$$
$$\alpha/R = v^* + (u^*)^{-2}$$
$$\beta/R = 1 + u^* + (\lambda/R)(u^*)^{-1}$$
$$\gamma/R = 1 + v^* - (\lambda/R)(v^*)^{-1}$$
$$\delta/R = u^* + (v^*)^{-2}$$
$$\varepsilon/R = 0.5(\lambda/R)(u^*)^{-1}(v^*)^{-1}(u^* - v^*)(2 + (u^* + v^*) - [3 + u^* + v^* + (u^*)^{-1} + (v^*)^{-1} + u^*v^*]$$
$$\mu/R = 2(1 - u^*v^*)[2(u^*)^2 + 2(v^*)^2 + u^*v^*(u^* + v^*)]/u^*v^*[2u^* + 2v^* + (u^*)^2 + (v^*)^2]$$

Notice that the equilibrium (u^*, v^*) is degenerate ($\lambda = 0$) precisely at the points where we already know tht only real eigenvalues can occur, namely where $d = 0$ ($u^*v^* = 1$) and on $\mathrm{Fix}(\pi_0)$ ($u^* = v^*$).

It remains to examine the higher order coefficient, which determines whether the Hopf equilibrium is a vague attractor or repellor. Marsden and McCracken (1976, p 126) have reduced this normal form computation to a formula that can be here applied with the uv coordinate system because of equation (16). Following Akin (1983) we denote by MARMC the number whose sign we wish to determine.

Proposition 3

When $u^*v^* \neq 1$ and $u^* \neq v^*$ and the parameter choices are given by *Table 2* then:

$$\text{MARMC} = \frac{3\pi}{4|\lambda/R|} \times \frac{\tilde{M}}{(u^*)^3(v^*)^3(u^*v^* - 1)}$$

$$\tilde{M} = 2(1 - u^*v^*)[(u^*)^3 + (v^*)^3] + (u^*)^2(v^*)^2[(u^*)^2 + (v^*)^2 + 2u^* + 2v^*)] \tag{26}$$

Furthermore, $\tilde{M} > 0$ if $u^*v^* \leq 1$ or if u^* and $v^* \geq 2$. For any $u^* < 2$, $\tilde{M}(u^*, v^*) = 0$ has a unique positive root in v^* and $\tilde{M}$ is negative when v^* is larger than this root. As $\tilde{M}(u^*, v^*) = \tilde{M}(v^*, u^*)$ the symmetric statement holds for any $v^* < 2$.

Proof

MARMC is independent of the choice of positive multiplicative constant and so we again normalize by fixing $R = 1$ (and again drop the asterisks). Because $\partial^2 F/\partial v^2$, $\partial^3 F/\partial u^3$, $\partial^2 G/\partial u^2$, $\partial^3 G/\partial v^3$ are identically zero, eight of the ten terms of the formula vanish and we are left with:

$$\frac{4|\lambda|}{3\pi}\text{MARMC} = \frac{1}{\lambda}\left(-\frac{\partial^2 F}{\partial u^2}\frac{\partial^2 F}{\partial u\,\partial v} + \frac{\partial^2 G}{\partial v^2}\frac{\partial^2 G}{\partial u\,\partial v}\right)$$

$$= \frac{1}{\lambda}[-2(\alpha - v)(\beta - 1 - 2u) + 2(\delta - u)(\gamma - 1 - 2v)]$$

Applying the *Table 2* equations gives:

$$\frac{4|\lambda|}{3\pi}\text{MARMC} = \frac{2}{\lambda}[v^{-2}(-\lambda v^{-1} - v) - u^{-2}(\lambda u^{-1} - u)]$$

$$= -\frac{2}{u^3v^3\lambda}[\lambda(u^3 + v^3) + u^2v^2(u - v)]$$

$$= -\frac{1}{u^3v^3(1 - uv)}[2(1 - uv)(u^3 + v^3) + u^2v^2(u^2 + v^2 + 2u + 2v)]$$

as in equation (26).

If $uv \leq 1$, then $\tilde{M}$ is clearly positive. If $u, v \geq 2$, then:

$$2uv(u^3 + v^3) \leq uv(u^3v + uv^3) = u^2v^2(u^2 + v^2)$$

and so the two negative terms in the expansion of $\tilde{M}$ are dominated by the next two terms.

Collecting terms according to powers of v:

$$\tilde{M} = u(u-2)v^4 + 2(1+u^2)v^3 + u^3(u+2)v^2 - 2u^4v + 2u^3$$
$$= v^3[u(u-2)v + 2(1+u^2)] + [u^3(u+2)v^2 - 2u^4 v + 2u^3]$$

When $0 < u < 2$ the coefficient of v^4 is negative and so $\tilde{M} < 0$ for v large. To prove that $\tilde{M} = 0$ has a unique positive root notice first that the latter quadratic portion is always positive because the discriminant:

$$(-2u^4)^2 - 4(2u^3)u^3(u+2) = 4u^8 - 8u^7 - 16u^6$$
$$< 8u^7 - 8u^7 - 16u^6 < 0$$

Also, the two higher order terms are positive for $0 < v < \tilde{v} \equiv 2(1+u^2)/u(2-u)$. The proof is completed by showing that for $v > \tilde{v}$, $\partial\tilde{M}/\partial v < 0$. Because the bracketed coefficient of v^3 is negative in this range:

$$\frac{\partial\tilde{M}}{\partial v} \leq v^3[u(u-2)] + [2u^3(u+2)v - 2u^4]$$
$$< u(u-2)\tilde{v}^2 v + 2u^3(u+2)v$$

and it suffices to show that $2u^3(u+2) \leq u(2-u)\tilde{v}^2$. Substituting the definition of $\tilde{v}$, a bit of manipulation reduces this to:

$$2u^4 \leq u^6 + 4u^2 + 2$$

Now, if $u \leq 1$, then $2u^4 < 4u^2$, while if $1 \leq u < 2$:

$$2u^4 \leq u^4 + 2^2u^2 \leq u^6 + 4u^2$$ QED

The cycles resulting from Hopf bifurcation (in the π_+ symmetric family) are locally stable in the uv plane when MARMC is negative, i.e., when either $M^*v^* < 1$ or $\tilde{M}(u^*,v^*) < 0$. Stability in the original three-dimensional system requires that the third eigenvalue be negative as well, and so $u^*v^* > 1$. Thus, limit cycles occur precisely from points (u^*,v^*) with $\tilde{M}(u^*,v^*) < 0$. For example, $(u^*,v^*) = (1,5)$.

References

Akin, E. (1979), The geometry of population genetics, *Lecture Notes in Biomathematics*, No. 31 (Springer-Verlag, Berlin).

Akin, E. (1982), Cycling in simple genetic systems, *J. Math. Biology*, **13**, 305–324.

Akin, E. (1983), *Hopf Bifurcation in the Two Locus Genetic Model*, Memoir No. 284 (Amer. Math. Soc., Providence, RI).

Antonelli, P. and Strobeck, C. (1977), The geometry of random drift I. Stochastic distance and diffusion, *Adv. Appl. Prob.*, **9**, 238–249.

Ewens, W. (1969), With additive fitness, the mean fitness increases, *Nature*, **221**, 1076.

Hastings, A. (1981), Stable cycling in discrete-time genetic models, *Proc. Natl. Acad. Sci. USA*, **78**, 7224–7225.

Karlin, S. and Feldman, M. (1970), Linkage and selection: Two locus symmetric viability models, *Theoret. Pop. Biology*, **1**, 39–71.

Kun, L. and Lyubich, Yu. (1979), Convergence to equilibrium under the action of additive selection in a multilocus multiallelic population, *Soviet Math. Dokl.* (AMS translations), **6**, 1380–1382.

Losert, V. and Akin, E. (1983), Dynamics of games and genes: Discrete versus continuous time, *J. Math. Biology*, **17**, 241–251.

Lyubich, Yu., Maistrovskii, G., and Ol'khovskii, Yu. (1980), Selection-induced convergence to equilibrium in a single-locus autosomal population, *Problems Inform. Transmission*, **16**, 66–75.

Marsden, J. and McCracken, M. (1976), *The Hopf Bifurcation and its Applications* (Springer-Verlag, Berlin).

Moran, P. (1964), On the nonexistence of adaptive topographies, *Ann. Human Genet.*, **27**, 383–393.

Shahshahani, S. (1979), *A New Mathematical Framework for the Study of Linkage and Selection*, Memoir No. 211 (Amer. Math. Soc., Providence, RI).

Traveling Fronts in Parabolic and Hyperbolic Equations

K. P. Hadeler
Lehrstuhl für Biomathematik, Universität Tübingen, Auf der Morgenstelle 10, 7400 Tübingen, Federal Republic of Germany

1. Introduction

Consider a reaction–diffusion equation

$$u_t = u_{xx} + f(u) \tag{1}$$

on the real line. For definiteness assume $f \in C^1(\mathbb{R})$, $|f(u)| \to \infty$ for $|u| \to \infty$, and that f has only finitely many zeros at $u_1 > u_2 > \cdots > u_n$, whereby $f'(u_i) \neq 0$ for $i = 1, \ldots, n$. A traveling front is a solution of equation (1) of the form

$$u(t,x) = \varphi(x - ct) \tag{2}$$

which has limits at $x = \pm\infty$. The function φ of one variable describes the shape of the front, and c is the speed of propagation. For such solution the function φ satisfies an ordinary differential equation

$$\varphi'' + c\varphi' + f(\varphi) = 0 \tag{3}$$

and side conditions

$$\varphi(-\infty) = u_- \qquad \varphi(+\infty) = u_+ \tag{4}$$

where u_-, u_+ are zeros of f.

Again, write $\varphi = u$ and $\varphi' = v$. Then equation (3) assumes the form

$$\dot{u} = v \qquad \dot{v} = -cv - f(u) \tag{5}$$

and the traveling front corresponds to a heteroclinic orbit of the system (5)

connecting two stationary points.

In most cases the traveling fronts are subject to further conditions of a more quantitative character. The typical requirements are:

(1) The heteroclinic orbit connects a saddle point to another saddle point.
(2) The heteroclinic orbit connects a saddle point at $(u_i,0)$ to a node at $(u_j,0)$ such that $u(t)$ stays between u_i and u_j for all $t \in \mathbf{R}$.

A mechanical interpretation is helpful. For $c = 0$ system (5) is a Hamiltonian system with the Hamiltonian function

$$H(u,v) = 0.5v^2 + F(u) \tag{6}$$

where

$$F(u) = \int_0^u f(s)\mathrm{d}s \tag{7}$$

describing a mass point running on a surface (curve) of potential energy defined by the function F. A zero u_i with $f'(u_i) < 0$ corresponds to a saddle point, a zero u_i with $f'(u_i) > 0$ is a center.

For $c \neq 0$ the motion of the mass point is subject to friction. Positive values of c correspond to positive (physical) friction, negative values of c must be interpreted as negative friction, which is physically somewhat unrealistic, but useful. Of course, this physical interpretation is a reformulation of the ideas of Conley (1978; see also Smoller, 1982).

The mechanical interpretation leads rather quickly to a complete picture of all traveling fronts.

To have a concrete situation at hand, assume $F(u) \to -\infty$ for $u \to +\infty$. Then u_1 is a maximum of potential energy and $(u_1,0)$ is a saddle point of system (5). By assumption, $u_3,u_5,u_7,\ldots$ correspond to saddle points, and $u_2,u_4,u_6,\ldots$ correspond to centers of the Hamiltonian system with $c = 0$.

To simplify notation we write u_i instead of $(u_i,0)$. For any u_i,i odd, we consider only that part of the unstable manifold that extends to smaller values of u. Consider especially u_1 and the asymptotic behavior of its unstable manifold. For large c – strong friction – the motion of the mass point starting at u_1 is overdamped, the trajectory converges to u_2 without ever leaving the strip $u_2 < u < u_1$. If c is decreased then one finds a minimal c_{12} for which the trajectory still connects to u_2 and does not leave $u_2 < u < u_1$. For $c < c_{12}$ the trajectory enters the strip $u < u_2$. There is a value $c_{12}^* \leq c_{12}$, $c_{12}^* > 0$ such that the motion near u_2 becomes oscillatory. Finally, there is a unique number c_{13} such that the unstable manifold of u_1 connects to u_3. Of course, the following two statements hold:

$$c_{13} \leq c_{12} \tag{8}$$

$$c_{13} \geq 0 \qquad \text{if and only if} \qquad F(u_3) \leq F(u_1) \tag{9}$$

Similarly, for the unstable manifold extending from u_3 toward u_4 one finds numbers c_{34}, c_{35} with

$$c_{35} \leq c_{34} \tag{10}$$

$$c_{35} \geq 0 \qquad \text{if and only if} \qquad F(u_5) \leq F(u_3) \tag{11}$$

Next we address the problem when the unstable manifold of u_1 enters the point u_5 for some c. If $c_{13} > c_{35}$ then the friction c_{13} is sufficient to connect u_1 to u_3 and also to keep the mass point starting from u_3 from running to u_5 (rather it ends up in u_4). Hence, some friction $c = c_{15}$ with $c_{35} < c_{15} < c_{13}$ will be sufficient to let the mass point run from u_1 to u_5. If, on the other hand, $c_{13} \leq c_{35}$ then for any $c < c_{13}$ the mass point starting at u_1 will pass u_5 because the friction is too low. In terms of phase portraits: the unstable manifold of u_1 cannot arrive at u_5 because the unstable manifold of u_3 separates it from u_5. Hence we find

$$c_{15} \text{ exists if and only if } c_{13} > c_{35} \tag{12}$$

These arguments can be extended to an arbitrary number of saddle points. Note that each saddle point is connected to its immediate successor, but not necessarily to other saddle points, i.e., c_{13} exists but c_{15} may not exist. Also, the existence of c_{17} does not imply the existence of c_{15}. A recursive characterization can be formulated as follows (all subscripts are odd).

Assume c_{1i} exists for some $i > 1$, i odd. Then

$$c_{1,i+2} \text{exists if and only if } c_{i,i+2} < c_{1i} \tag{13}$$

Assume c_{1i} exists for some $i > 1$, i odd, and $c_{1,i+2}, \ldots, c_{1,j}$ do not exist for some $j > i+1$. Then

$$c_{1,j+2} \text{ exists if and only if } c_{i,j+2} \text{ exists and } c_{i,j+2} < c_{1i} \tag{14}$$

First we prove condition (13). Assume $c_{i,i+2} < c_{1i}$. For $c = c_{1i}$ the unstable manifold of u_1 connects to u_i and the unstable manifold of u_i connects to u_{i+1}. For c slightly less than c_{1i} the unstable manifold of u_1 intersects the axis $v = 0$ between u_{i+2} and u_{i+1}. If c is further decreased, then at some $c = c_{1,i+2}$ this unstable manifold connects to u_{i+2}. If, on the other hand, $c_{1,i+2}$ exists then a trajectory connects u_1 and u_{i+2} and consequently $c_{i,i+2}$ is less than c_{1i}.

The proof of condition (14) is essentially the same. Assume $c_{i,i+2} < c_{1i}$. For $c = c_{1i}$ the unstable manifold of u_1 connects to u_i and the unstable manifold of u_i connects to u_{j+1}. For c slightly less than c_{1i} the unstable manifold of u_i intersects the axis $v = 0$ between u_{j+2} and u_{j+1}. If c is further decreased, then at some $c = c_{1,j+2}$ this unstable manifold connects to u_{j+2} before any other saddle point, in particular u_i, can be connected to u_{j+2}. If, on the other hand, $c_{1,j+2}$ exists then a trajectory connects u_1 and u_{j+2} and, consequently, $c_{i,j+2} < c_{1i}$.

Consider the special case where

$$f(0) = f(1) = 0 \qquad f(u) > 0 \text{ for } 0 < u < 1 \qquad f'(0)\,f'(1) < 0$$

and restrict to solutions confined to $0 \le u \le 1$.

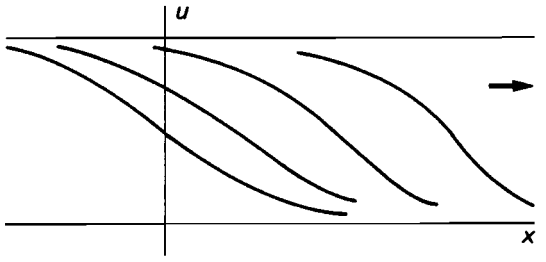
A typical equation is

$$u_t = 0.5u_{xx} + u(1-u) \tag{15}$$

This equation has two quite different interpretations. Fisher (1937) and Kolmogorov *et al.* (1937) consider a population that is locally governed by a Verhulst equation $\dot{u} = 0.5u(1-u)$ and that also diffuses. Thus, $u(t,x)$ is the local population density at the space point x at time t. The advance of the traveling front results from joint effects of the nonlinear iteration f and diffusion. If one considers the initial datum

$$u(0,x) = \begin{cases} 1 & x < 0 \\ 0 & x \ge 0 \end{cases} \tag{16}$$

and looks at diffusion and interaction as alternating processes, then evolution into a traveling front is intuitively clear:



The traveling front describes the progress of the bulk of the population.

McKean (1975) has defined a branching process with Brownian motion. At a given time there are n particles located at positions $x_1(t), \ldots, x_n(t)$.

Each particle performs Brownian motion. Independently, the particles undergo a branching process. At a given time t let $u(t,x)$ be the probability that not all particles are located to the left of x. Thus, if there is originally a single particle at $x = 0$, then $u(0,x)$ is given by condition (16). For $t > 0$ the profile of u should move to the right. McKean showed that, for the case where each particle has exactly two daughters, the function u satisfies equation (15). Thus, in McKean's interpretation the equation describes a probability for the position of the most advanced particle.

For more general branching processes McKean derives equations of the form (1) with a general nonlinearity f.

Mostly, the following two problems have been investigated.

(1) Type I, positive f

$$\begin{aligned} &f(0) = f(1) = 0 \\ &f(u) > 0 \text{ for } 0 < u < 1 \\ &f'(0)f'(1) < 0 \end{aligned} \tag{17}$$

(2) Type II, threshold f

$$\begin{aligned} &f(0) = f(1) = 0 \qquad f(\alpha) = 0 \text{ for some } \alpha \in (0,1) \\ &f(u) > 0 \text{ for } 0 < u < \alpha \qquad f(u) > 0 \text{ for } \alpha < u < 1 \\ &f'(0)f'(\alpha)f'(1) < 0 \end{aligned} \tag{18}$$

If we restrict attention to fronts with

$$0 \le \varphi(x) \le 1 \tag{19}$$

$$\varphi(-\infty) = 1 \qquad \varphi(+\infty) = 0 \tag{20}$$

then the following is true.

In Type I there is a number $c_0 > 0$ such that for each $c \in [c_0,\infty)$ there is a traveling front with speed c. This traveling front is unique up to translation.

In terms of the phase portrait of system (5), the traveling front corresponds to a heteroclinic orbit connecting the saddle point (1,0) to the node (0,0). For values $c \in (0,c_0)$ the unstable manifold of (1,0) is still connected to (0,0), but the trajectory does not stay in $0 \le u \le 1$. In physical terms: for each friction $c \in [c_0,\infty)$ there is a motion of the mass point starting at the saddle point (1,0) and ending in (0,0) such that the mass point stays in [0,1]. For $c \in (0,c_0)$ the friction is not strong enough to keep the mass point from overshooting.

In Type II there is a single number c_0 such that there is a traveling front with speed c_0. This traveling front is unique up to translation.

In terms of the phase portrait of system (5), the traveling front is a heteroclinic orbit connecting the two saddle points (1,0) and (0,0). In physical terms: there is only one value of the friction parameter c that allows a motion starting at the saddle point (1,0) and ending at (0,0). The number c_0 is

positive if and only if

$$\int_0^1 f(u)\mathrm{d}u > 0 \tag{21}$$

In Hadeler (1981) the minimal (Type I) resp. unique (Type II) speed c_0 has been characterized by minimax and maximin principles over appropriate sets of functions (see also Hadeler and Rothe, 1975).

2. Diffusion Rate Depending on the Unknown Function

I have generalized (Hadeler, 1981, 1983) equation (1) to

$$m(u)u_t = [k(u)u_x]_x + f(u) \tag{22}$$

Here f is a production term as before, k can be interpreted as a density-dependent diffusion rate, and m as a density-dependent capacity. Of course, the parameter function m does not make much sense from the viewpoint of physics or ecology, but it is very helpful in the following hyperbolic problem.

I assume that m, k, f are continuously differentiable functions and that m, k are strictly positive in [0,1] and f is of Type I or II above.

The ansatz for a traveling front $u(t,x) = \varphi(x - ct)$ leads to

$$[k(\varphi)\varphi']' + cm(\varphi)\varphi' + f(\varphi) = 0 \tag{23}$$

Although the partial differential equation (22) cannot be reduced to a problem of type (1), such a reduction is possible for equation (23) (see Hadeler, 1983; Engler, 1985). Define again $\varphi = u$, $\varphi' = v$. Then equation (23) reads

$$\begin{aligned} \dot{u} &= v \\ [k(u)v]^{\cdot} + cm(u)v &+ f(u) = 0 \end{aligned} \tag{24}$$

Define

$$H = \int_0^1 m(s)\mathrm{d}s \qquad h(u) = \frac{1}{H}\int_0^u m(s)\mathrm{d}s \tag{25}$$

$$\tilde{u} = h(u) \tag{26}$$

$$\tilde{f}(\tilde{u}) = \frac{1}{H} f(u)k(u) \qquad u = h^{-1}(\tilde{u}) \tag{27}$$

$$\tau = \int_0^t \frac{m[u(s)]}{k[u(s)]}\,ds \tag{28}$$

Then the functions $u(\tau)$, $v(\tau)$ satisfy

$$\begin{aligned} \frac{d\tilde{u}}{d\tilde{\tau}} &= \tilde{v} \\ \frac{d\tilde{v}}{d\tau} &= -c\tilde{v} - \tilde{f}(\tilde{u}) \end{aligned} \tag{29}$$

Hence the functions $\tilde{u}$, $\tilde{v}$ satisfy equations of the form (5). Here $\tilde{f}$ is of Type I (or II, respectively) if and only if f is of Type I (or II, respectively). Notice that the parameter c is the same in equations (23) and (29). Hence, there is a complete description of the traveling fronts for equation (22).

If in equation (22) the function f is of Type I, then there is a number c_0 depending on m, k, f such that equation (22) has a traveling front solution, unique up to translation, for every $c \geq c_0$.

If in equation (22) the function f is of Type II, then there is one and only one number c_0 depending on m, k, f such that equation (22) has a traveling front solution, which is unique up to translation. From equations (21) and (23) it follows immediately that c_0 is positive if and only if

$$\int_0^1 f(u)k(u)m(u)du > 0 \tag{30}$$

3. Hyperbolic Traveling Fronts

Dunbar and Othmer (1986) have introduced a branching process with Poisson migration. Again, at time t the population consists of finitely many particles located at $x_1(t), \ldots, x_n(t)$. The particles move according to the following laws. The particles move with speed $+v_0$ or $-v_0$. They switch between these states with exponential holding time. Also, independently of the motion, the particles are subject to a branching process where each particle splits into two daughters. It is assumed that the speed of a daughter is independent of the speed of the mother ($\pm v_0$ each with probability 0.5).

Again, let $u(t,x)$ be the probability that at time t not all particles are to the left of x. Dunbar and Othmer (1986) show that u satisfies a hyperbolic equation of the form

$$\varepsilon^2 u_{tt} + g(u)u_t = [k(u)u_x]_x + f(u) \tag{31}$$

In the actual application k is a constant and g, f are polynomials in u.

I assume that f is of Type I defined in conditions (17) and that g, k are continuously differentiable and positive.

The ansatz $u(t,x) = \varphi(x - ct)$, and then $\varphi = u$, $\varphi' = v$ leads to

$$\dot{u} = v$$

$$\varepsilon^2 c^2 \dot{v} + cg(u)v = [k(u)v] + f(u) \tag{32}$$

Again, this problem can be reduced by an appropriate substitution. Define

$$H = \int_0^1 g(s)ds \qquad G(u) = \frac{1}{H} \int_0^u g(s)ds \tag{33}$$

and

$$\left.\begin{aligned} \tilde{k}(\tilde{u}) &= k(u) \\ \tilde{f}(\tilde{u}) &= \frac{1}{H} \frac{f(u)}{g(u)} \end{aligned}\right\} \qquad u = G^{-1}(\tilde{u}) \tag{34}$$

Then the variables

$$\begin{aligned} \tilde{u} &= G(u) \\ \tilde{v} &= \frac{1}{H}[k(u) - \varepsilon^2 c^2]v \end{aligned} \tag{35}$$

satisfy

$$\begin{aligned} \dot{\tilde{u}} &= \tilde{v} \\ \dot{\tilde{v}} &= -c\tilde{v} - [\tilde{k}(\tilde{u}) - \varepsilon^2 c^2]\tilde{f}(\tilde{u}) \end{aligned} \tag{36}$$

This system has the same properties as system (5), provided the factor $\tilde{k}(\tilde{u}) - \varepsilon^2 c^2$ is positive.

One can show the following (Hadeler, 1985a):

(1) If $k \equiv 1$, then for problem (31) there is a bounded interval $[c_0, 1/\varepsilon)$ of speeds. The minimal speed is given by

$$c_0 = \frac{c_F}{(1 + \varepsilon^2 c_F^2)^{0.5}} \tag{37}$$

(2) Let k be a strictly increasing function. Then there is a number $\varepsilon^* > 0$ such that for each $\varepsilon \in (0,\varepsilon^*)$ there is a continuum of speeds of traveling fronts $[c_h(\varepsilon), \alpha_1/\varepsilon)$, where $\alpha_1 = [k(0)]^{0.5}$. For $\varepsilon > \varepsilon^*$ there are no proper traveling fronts.

(3) Let k be a strictly decreasing function. Then there is a number $\varepsilon^* > 0$ such that for each number $\varepsilon \in (0,\varepsilon^*)$ there is a continuum of speeds of traveling fronts $[c_h(\varepsilon), \alpha_1/\varepsilon)$ where $\alpha_1 = [k(1)]^{0.5}$. For $\varepsilon > \varepsilon^*$ the front splits into two layers traveling at differing speeds.

4. Traveling Fronts and the Hypercycle

Recently a relation between the hypercycle equations of Eigen and Schuster and traveling fronts was found (Hadeler, 1986). The hypercycle is a system of ordinary differential equations

$$\dot{u}_i = u_i u_{i-1} - \sum_{j=1}^{n} u_j u_{j-1} \times u_i \tag{38}$$

where the variables are taken in cyclic order. A continuous analogon is the equation

$$u_t(t,x) = u(t,x)u(t,x-r) - \int_0^1 u(t,y)u(t,y-r)\mathrm{d}y \times u(t,x) \tag{39}$$

where $x \in \mathbb{R} \bmod l$ is a variable on a circle of circumference l, and $0 < r < l$. Hence $u(t,x+l) = u(t,x)$.

The ansatz for a traveling wave $u(t,x) = \varphi(\xi)$, $\xi = x - ct$, $\varphi(\xi) = \varphi(\xi + l)$, leads to the equation

$$\varphi'(\xi) = -\frac{1}{c}\varphi(\xi)[\varphi(\xi - r) - H] \tag{40}$$

where

$$H = \int_0^1 \varphi(y)\varphi(y-r)\mathrm{d}y \tag{41}$$

Then the transformation

$$\varphi = H(1 + v)$$

gives

$$v'(\xi) = -\frac{H}{c}[1 + v(\xi)]v(\xi - r) \tag{42}$$

Finally, putting $\xi = rt, v(\xi) = u(t)$ gives

$$\dot{u}(t) = -\alpha[1 + u(t)]u(t - 1) \tag{43}$$

where

$$\alpha = H r / c \tag{44}$$

and the required period for the function u is

$$p = l / r \tag{45}$$

Equation (43) is Wright's equation. From known results on this equation one can derive the following existence result. Choose $l > 0$, then for any delay $r = l/p < l/4$, the continuous hypercycle (39) has a traveling wave solution with speed $c = 1/(\alpha p)$, where α is the parameter in (43) for which a periodic solution of period $p \in (4, \infty)$ occurs.

References

Conley, C. (1978), Isolated invariant sets and the Morse index, *C.B.M.S. Notes*, **38** (Amer. Math. Soc., Providence).

Dunbar, S. and Othmer, H. (1986), On a nonlinear hyperbolic equation describing transmission lines, cell movement, and branching random walks, in H. G. Othmer (Ed), *Nonlinear Oscillations in Biology and Chemistry*, Lecture Notes in Biomathematics (Springer-Verlag, Berlin, Heidelberg, New York, Tokyo).

Engler, H. (1985), Relations between traveling wave solutions of quasilinear parabolic equations, *Proc. Amer. Math. Soc.*, **93**, 297–302.

Fisher, R. A. (1937), The advance of advantageous genes, *Ann. of Eugenics*, **7**, 355–369.

Hadeler, K. P. (1981), Travelling fronts and free boundary value problems, in J. Albrecht, L. Collatz, and K. H. Hoffmann (Eds), *Numerical Treatment of Free Boundary Value Problems*, pp 90–107, Oberwolfach Conference 1980 (Birkhäuser Verlag).

Hadeler, K. P. (1983), Free boundary problems in biological models, in A. Fasano and M. Primicerio (Eds), *Free Boundary Problems: Theory and Applications*, Vol. II, pp 664–671, Montecatini Conference, 1981 (Pitman).

Hadeler, K. P. (1986), The hypercycle, traveling waves, and Wright's equatins, *J. Math. Biol.* (in press).

Hadeler, K. P. (1987), Hyperbolic traveling fronts, *Proc. Edinb. Math. Soc.* (in press).

Hadeler, K. P. and Rothe, F. (1975), Traveling fronts in nonlinear diffusion equations, *J. Math. Biol.*, **2**, 251–263.

Kolmogorov, A., Petrovskij, I., and Piskunov, N. (1937), Etude de l'equation de la diffusion avec croissance de la quantité de la matière et son application à une problème biologique, *Bull. Univ. Moscou, Ser. Int., Sec. A.*, **1**(6), 1–25.

McKean, H. P. (1975), Application of Brownian motion to the equation of Kolmogorov–Petrovskij–Piskunov, *Comm. Pure Appl. Math.*, **28**, 323–331.

Smoller, J. (1982), *Shock Waves and Reaction Diffusion Equations* (Springer-Verlag, Berlin, Heidelberg, New York, Tokyo).

Competitive Exclusion by Zip Bifurcation

M. Farkas
Technical University of Budapest, Mathematics Department, Faculty of Medical Engineering, XI, Müegyetem Rakpart 3, Budapest, Hungary

1. Introduction

A model that describes the competition of two predator species for a single regenerating prey species was introduced by Hsu *et al.* (1978 a,b; see also Koch, 1974 a,b) and has been studied since then by several workers, e.g., Butler (1983), Keener (1983), Smith (1982), and Wilken (1982). In this model of a three-dimensional system of ordinary differential equations the prey population is assumed to have a logistic growth rate in the absence of predators, and the predator populations are assumed to obey a Holling-type functional response (Michaelis–Menten kinetics). Butler (1983) has shown that most of the results concerning the model of Hsu *et al.* can be achieved for a whole class of two-predator–one-prey models whose common feature is that the prey's growth rate and the predators' functional response are arbitrary functions satisfying certain natural conditions.

We have studied the model of Hsu *et al.* under the special assumption that the values of a certain threshold parameter of the predator species are equal at the two predators. This assumption made possible the identification of one of the predators of equal thresholds as an "r-strategist" and the other as a "K-strategist". Here "r-strategist" means a predator whose ratio of maximal birth rate to death rate is high but which needs a large amount of food to increase birth rate, while a "K-strategist" means a predator with a relatively low ratio of maximal birth rate to death rate but with the ability of keeping the birth rate relatively high even if only small amounts of food are available. Farkas (1984, 1985) has introduced the concept of zip bifurcation to denote the following phenomenon. At low values of the carrying capacity K of the ecosystem with respect to the prey a line of equilibria, which is an attractor of the system, represents stable coexistence of the three species. If K is increased then the equilibria are continuously destabilized, starting with those which represent the dominance of the K-strategist over the r-strategist. Above a certain value of K the system has no more stable equilibria representing coexistence; however, a stable limit cycle representing the oscillating coexistence of the r-strategist and the prey remains.

In this paper we shall show that the phenomenon of zip bifurcation is general, i.e., it is present in all the systems satisfying Butler's conditions. A new characteristic of the relative fitness of the two predator populations will

also be introduced which has the important role of determining the "direction of the zip", i.e., the outcome of the competition between an r-strategist and a K-strategist when K is increased.

As a consequence of our assumption all the models considered become structurally unstable. Now, according to May (1981) "structurally unstable models have no place in biology". Accepting this maxim, we do not recommend these models under our assumption for describing real particular ecosystems. Their study, nevertheless, may turn out to be useful for two reasons. First, our assumption creates an abstract pure situation in which two predators of equal prey thresholds compete, one achieving this threshold by being an r-strategist and the other by being a K-strategist. Secondly, bifurcation of a stable periodic solution representing coexistence can be proved using our assumption and then moving the respective thresholds away from their common value (see Keener, 1983, and Smith, 1982).

In Section 2 the model will be introduced with the conditions imposed. In Section 3 the stability of the equilibria will be studied and the occurrence of zip bifurcation will be shown as a consequence of the variation of the carrying capacity. In Section 4 examples will be given.

2. A General Class of Two-Predators–One-Prey Models

We denote the quantities of prey and the ith predator at time t by $S(t)$ and $x_i(t)$ respectively ($i = 1, 2$). We assume that the per capita growth rate of prey in the absence of predators is $\gamma g(S,K)$ where γ is a positive constant (in fact the maximal growth rate of the prey), $K > 0$ is the carrying capacity of the environment with respect to the prey, the function g satisfies the conditions $g \in C^2((0,\infty) \times (0,\infty), \mathbf{R})$, $g \in C^0([0,\infty) \times (0,\infty), \mathbf{R})$,

$$g(0,K) = 1 \qquad g'_S(S,K) < 0 < g''_{SK}(S,K) \qquad S \geq 0, K > 0 \tag{1}$$

$$\lim_{K \to \infty} g'_S(S,K) = 0 \tag{2}$$

uniformly in $[\delta, S_0]$ for any $0 < \delta < S_0$, and the (possibly) improper integral $\int_0^{S_0} g'_S(S,K)\mathrm{d}S$ is uniformly convergent in $[K_0, \infty)$ for any $K_0 > 0$:

$$(K - S)g(S,K) > 0, \qquad S \geq 0 \qquad K > 0 \qquad S \neq K \tag{3}$$

We assume further that the death rate d_i of the ith predator is constant ($d_i > 0$) and that the per capita birth rate of the same predator is $p(S, a_i)$, where a_i is a positive constant ($i = 1, 2$) and the function p satisfies the following conditions:

$$p \in C^1((0,\infty) \times (0,\infty), \mathbf{R}) \qquad p \in C^0([0,\infty) \times (0,\infty), \mathbf{R})$$

$$p(0,\alpha) = 0 \qquad p'_S(S,\alpha) > 0 \qquad S > 0 \qquad \alpha > 0 \tag{4}$$

$$p'_S(S,\alpha) < p(S,\alpha)/S \qquad S > 0 \qquad \alpha > 0 \tag{5}$$

$$p'_\alpha(S,\alpha) < 0 \qquad S > 0 \qquad \alpha > 0 \tag{6}$$

We finally assume that the presence of predators decreases the growth rate of prey exactly by the amount equal to the birth rate of the respective predator.

Under these assumptions the dynamics of the ecosystem consisting of the three species is described by the system of differential equations

$$\begin{aligned} \dot{S} &= \gamma S g(S,K) - x_1 p(S,\alpha_1) - x_2 p(S,\alpha_2) \\ \dot{x}_1 &= x_1 p(S,\alpha_1) - d_1 x_1 \\ \dot{x}_2 &= x_2 p(S,\alpha_2) - d_2 x_2 \end{aligned} \tag{7}$$

where the dot denotes differentiation with respect to time t.

At this point some remarks about conditions (1)–(6) are appropriate. These conditions coincide with those of Butler (1983), apart from a genericity condition made by Butler which is unnecessary in our study. Conditions (1) mean that the highest specific growth rate of prey is achieved at $S = 0$, $x_1 = x_2 = 0$, and it is $\gamma > 0$; the growth rate decreases if the quantity of prey increases, and the rate of decrease in growth rate g'_S is negative and an increasing function of the carrying capacity K, i.e., the effect of the increase in prey diminishes with increase in K. Relation (2) means that for very high values of K changes in the quantity of prey have a negligible effect on the growth rate. The inequality (3) means that (in the absence of predators) the growth rate of prey is positive if S is below the carrying capacity K and negative if S is above it. It is easy to see that conditions (1)–(2) imply that

$$\lim_{K \to \infty} g(S,K) = 1 \qquad S \geq 0 \tag{8}$$

Conditions (4) mean that the per capita birth rate p of the predators (also called the "predation rate" or the "functional response") is zero in the absence of prey and is an increasing function of the quantity of prey. Condition (5) is a "weak concavity" condition, sometimes called Krasonselskij's condition. If p is a strictly concave function of S (for any $\alpha > 0$), (5) is implied with the possible exception of isolated points where it holds with an equality sign. Inequality (6) throws light on the role of the parameter α; the birth rate of the predator is a decreasing function of α, i.e., the higher is the value of α the more food is needed to maintain the same birth rate of the specific predator. [In the original model of Hsu *et al.* (1978b) α is the "half-saturation constant".] The conditions imply that $p(S,\alpha)$, S, and α are all greater than zero. In the case where p is a bounded function for fixed

$a > 0$,

$$m_i = \sup_{S>0} p(S,a_i)$$

is the "maximal birth rate" of the ith predator. Clearly,

$$\lim_{S\to\infty} p(S,a_i) = m_i \qquad \text{if } p \text{ is bounded}$$
$$= \infty \qquad \text{if } p \text{ is not bounded} \tag{9}$$

In the second and third terms on the right-hand side of the differential equation describing the growth of the prey population some constants (called "yield factors") used to appear in a realistic model. However, these yield factors can be transformed out of the system without loss of generality and they do not affect the qualitative behavior. Therefore we have chosen not to introduce them at all.

Omitting the analysis of the less interesting case in which a_1 and a_2 are equal, we shall assume $a_1 > a_2 > 0$, i.e.,

$$p(S,a_1) < p(S,a_2) \qquad \text{for all} \qquad S > 0 \tag{10}$$

According to this condition, at any given level of prey quantity the birth rate of predator 2 is higher than that of predator 1 or, in other words, predator 1 needs a higher quantity of prey to achieve the same birth rate as predator 2. Now, if d_1 is greater than d_2, (10) implies $p(S,a_1) - d_1 < p(S,a_2) - d_2$, i.e., the net growth rate of predator 2 is higher than that of predator 1. We can prove that in this case predator 2 out-competes predator 1 provided that the conditions for the survival of the former hold (see later). We assume from now on that

$$d_1 < d_2 \tag{11}$$

As a consequence, (10) does *not* imply that the net growth rate of predator 2 also exceeds that of predator 1.

Another important characteristic of the respective predator species is the prey threshold quantity $S = \lambda_1$, above which their growth rate is positive, i.e.,

$$p(\lambda_i,a_i) = d_i \qquad i = 1, 2$$

Obviously, the lower is λ_i the fitter is predator i. However, we shall assume (nongenerically) that $\lambda_1 = \lambda_2$, i.e., the two predator species have equal prey

thresholds although they achieve this by different means. Thus our assumption will be that there exists a $\lambda > 0$ such that

$$p(\lambda, a_i) = d_i \qquad i = 1, 2 \tag{12}$$

We note that, because of condition (4), equation (12) has one and only one solution λ if and only if either p is unbounded or m_i is greater than d_i. The real content of (12) is that the two solutions for $i = 1, 2$ coincide.

The class of models under consideration will be divided into three subclasses according to *Definition 1*.

Definition 1

We say that the model (7) under conditions (1)–(6) and (10)–(12) is natural, artificial, and degenerate if

$$\frac{\partial}{\partial S}\left[\frac{p(S, a_2)}{p(S, a_1)}\right]_{S=\lambda} \begin{cases} < 0 \\ > 0 \\ = 0 \end{cases} \tag{13}$$

respectively.

The first inequality of (13) means that, by continuity, the ratio of the birth rates (which is, by (10), greater than unity) decreases in the neighborhood of $S = \lambda$, i.e., the advantage of species 2 over species 1 expressed by (10) decreases as the quantity of prey increases. This is what is usually expected to happen. The second inequality of (13) means that the same advantage is increasing. The importance of the point $S = \lambda$ will become clear in what follows.

Before turning to the study of the equilibria of system (7) we note that, obviously, the coordinate planes of S, x_1, x_2 space are invariant manifolds of the system, and that it can be proved by standard methods that all solutions with nonnegative initial conditions of the system are defined in $[0, \infty)$, are bounded, and remain nonnegative.

System (7) has the following equilibria: $Q_1 = (0,0,0)$, $Q_2 = (K,0,0)$, and the points on the straight line segment

$$\begin{aligned} L_K = \{(S, x_1, x_2) : p(\lambda, a_1)x_1 + p(\lambda, a_2)x_2 = \gamma\lambda g(\lambda, K), \\ S = \lambda,\ x_1 \geq 0,\ x_2 \geq 0\} \end{aligned} \tag{14}$$

It is easy to see by linearization that Q_1 is unstable, and Q_2 is asymptotically stable for $K < \lambda$ and unstable for $K > \lambda$. Actually, it is known (see Hsu *et al.*, 1978 a; Butler, 1983) that

$$K > \lambda \tag{15}$$

is a necessary condition for the survival of each predator. Therefore (15) will also be assumed in what follows. Note that, by (3), if K is less than λ then L_K is empty, and if $K = \lambda$ then its only point is the origin Q_1.

3. Coexistence and Extinction by Zip Bifurcation

In this section the stability of the set L_K is studied. The elements of the set are denoted by (λ,ξ_1,ξ_2), i.e., $(\lambda,\xi_1,\xi_2) \in L_K$.

Linearizing system (7) at an arbitrary point (λ,ξ_1,ξ_2) of L_K gives the characteristic polynomial of the linearized system as

$$D(\mu) = \mu[\mu^2 + \{\xi_1 p'_S(\lambda,a_1) + \xi_2 p'_S(\lambda,a_2) - \gamma' g(\lambda,K) - \gamma\lambda g'_S(\lambda,K)\}\mu + \xi_1 p(\lambda,a_1) p'_S(\lambda,a_1) + \xi_2 p(\lambda,a_2) p'_S(\lambda,a_2)]$$

Thus $\mu = 0$ is always an eigenvalue. The quadratic polynomial in square brackets is stable, i.e., the remaining two eigenvalues have negative real parts if and only if

$$\xi_1 p'_S(\lambda,a_1) + \xi_2 p'_S(\lambda,a_2) > \gamma[g(\lambda,K) + \lambda g'_S(\lambda,K)]$$

We rewrite the above inequality as follows:

$$\gamma\lambda g(\lambda,K) + \gamma\lambda^2 g'_S(\lambda,K) < [\lambda p'_S(\lambda,a_1) - p(\lambda,a_1)]\xi_1 + [\lambda p'_S(\lambda,a_2) - p(\lambda,a_2)]\,\xi_2 + p(\lambda,a_1)\xi_1 + p(\lambda,a_2)\,\xi_2$$

Taking into account the fact that (ξ_1,ξ_2) satisfies the equation in (14), we obtain

$$[p(\lambda,a_1) - \lambda p'_S(\lambda,a_1)]\xi_1 + [p(\lambda,a_2) - \lambda p'_S(\lambda,a_2)]\xi_2 < -\gamma\lambda^2 g'_S(\lambda K) \tag{16}$$

In view of condition (5) the left-hand side is positive for all $(\lambda,\xi_1,\xi_2) \in L_K$. In view of (1) and (2) the right-hand side is positive, and decreases and tends to zero for $K \to \infty$. Let us consider the straight line segment B_K given by (16) taken with an equals sign:

$$
\begin{aligned}
B_K = \{(S,x_1,x_2) : &[p(\lambda,a_1) - p'_S(\lambda,a_1)]x_1 \\
&+ [p(\lambda,a_2) - \lambda p'_S(\lambda,a_2)]x_2 \\
&= -\gamma\lambda^2 g'_S(\lambda,K),\ S = \lambda,\ x_1 \geq 0,\ x_2 \geq 0\}
\end{aligned}
\tag{17}
$$

As a preparation for *Theorem 1* we determine the point of intersection of the lines of L_K and B_K. Denoting the point of intersection by $[x_1(K),x_2(K)]$ we obtain

$$
x_1(K) = -\gamma \frac{-\lambda g'_S(\lambda,K)p(\lambda,a_2) + g(\lambda,K)[\lambda p'_S(\lambda,a_2) - p(\lambda,a_2)]}{p(\lambda,a_2)p'_S(\lambda,a_1) - p'_S(\lambda,a_2)p(\lambda,a_1)}
\tag{18}
$$

$$
x_2(K) = \gamma \frac{-\lambda g'_S(\lambda,K)p(\lambda,a_1) + g(\lambda,K)[\lambda p'_S(\lambda,a_1) - p(\lambda,a_1)]}{p(\lambda,a_2)p'_S(\lambda,a_1) - p'_S(\lambda,a_2)p(\lambda,a_1)}
$$

provided that the model is not degenerate, i.e., the denominator is different from zero.

Theorem 1

Assume that system (7) satisfies conditions (1)–(6) and (10)–(12), and that it is natural. Then there exist $\lambda < K_1 < K_2 < \infty$ such that for $K \in (\lambda,K_1)$ all points of segment L_K are stable in the Ljapunov sense, and L_K is an attractor of the system. For $K \in (K_2,\infty)$ the system has no stable equilibrium in the closed positive octant of S, x_1, x_2 space. For $K \in (K_1,K_2)$ the point $[\lambda,x_1(K),x_2(K)]$ divides L_K into two parts (one of which may be empty): the points of L_K to the left of this point are unstable, the points to the right are stable in the Ljapunov sense, and the part of L_K to the right of this point is an attractor of the system.

Proof

By assumption our model is natural, i.e., the denominator of both $x_1(K)$ and $x_2(K)$ is positive. The conditions imposed upon the functions g and p imply $x_1(\lambda) < 0$, $x_2(\lambda) > 0$, and

$$
\lim_{K\to\infty} x_1(K) > 0
$$

$$
\lim_{K\to\infty} x_2(K) < 0
$$

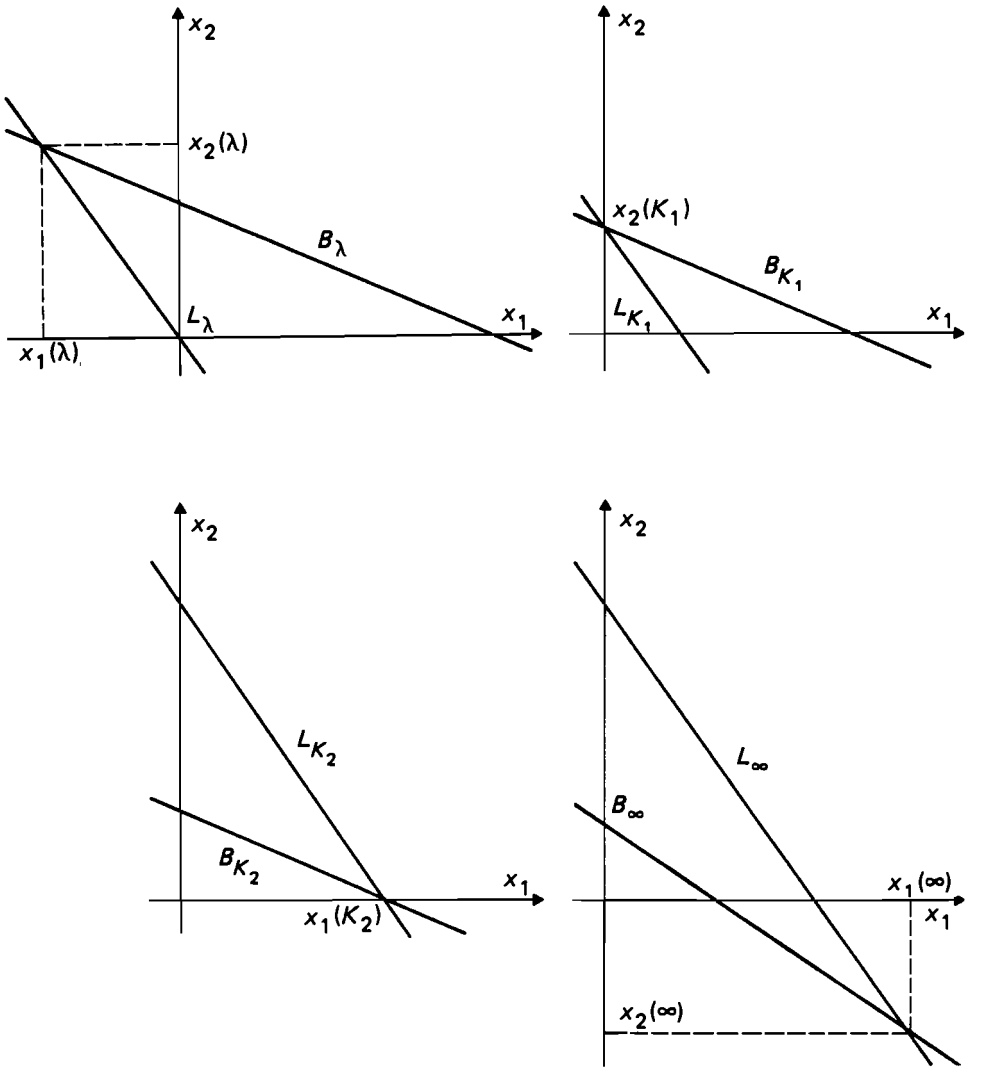


Figure 1. Points of intersection of the lines of B_K and L_K at extreme values of K in the case of a natural model: the slope of B_K is greater than the slope of L_K.

Thus, by continuity, for increasing K there exists a "first" $K_1 > \lambda$ at which $x_1(K_1) = 0$. However, since the line of L_K intersects axis x_2 at a point of positive coordinate for $K = K_1 > \lambda$, $x_2(K_1)$ is greater than zero. Similarly, for increasing K there exists a "last" $K_2 > K_1$ at which $x_2(K_2) = 0$ (*Figure 1*). This means that, for $K \in (\lambda, K_1)$, the lines of L_K and B_K intersect outside the positive quadrant and the set L_K is "below" the line B_K, i.e., condition (16) holds at each point of L_K. However, (16) implies that at these points the linearized system has two eigenvalues with negative real parts, i.e., each

point of L_K has a two-dimensional stable manifold. For $K \in (K_2, \infty)$ the set L_K is "above" the line of B_K, i.e., condition (16) with an opposite inequality sign holds at each point of L_K. This means that these points are unstable. If $K \in (K_1, K_2)$ either $[\lambda, x_1(K), x_2(K)] \in L_K$ (this is necessarily the case if $x_1(\cdot)$ is an increasing function and $x_2(\cdot)$ a decreasing function) or it is again to the left or to the right outside the positive octant. In the latter case one of the previous arguments applies. In the former case it divides L_K into two parts: in the left-hand part condition (16) with an inverted inequality sign holds, and in the right-hand part (16) is valid (*Figure 2*). From now on the proof coincides with the proof of the corresponding *Theorem 1* given by Farkas (1984).

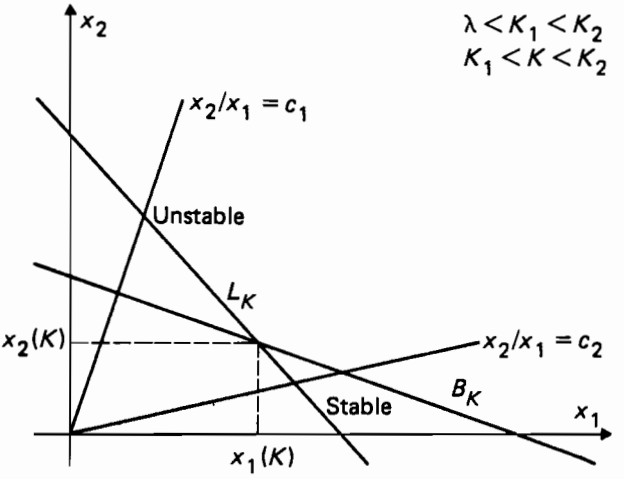


Figure 2. "Partly open zip" in a natural model with unstable and stable ratios C_1 and C_2, respectively, at level K.

Note that, if $g(\lambda, \cdot)$ is a nondecreasing function in $K \in (\lambda, \infty)$, i.e.,

$$g'_K(\lambda, K) \geq 0 \qquad K > \lambda \tag{19}$$

then, obviously, $x_1(\cdot)$ is a monotonic increasing function and $x_2(\cdot)$ is a monotonic decreasing function; as a consequence, if K is increased from K_1 to K_2 the point $[\lambda, x_1(K), x_2(K)]$ moves steadily along L_K from the left-hand end to the right-hand end while the segment L_K undergoes a parallel displacement "upwards". In the process the points left behind by $[\lambda, x_1(K), x_2(K)]$ become destabilized. We have called this phenomenon a *zip bifurcation*. In the more general case, when (19) does not hold, the movement of the point $[\lambda, x_1(K), x_2(K)]$ which separates the stable from the unstable equilibria is still continuous and yields, in the end, the same result, i.e., "the opening up of the zip"; however, in some parts of the interval (K_1, K_2) the movement may occur "backwards" (to the left) with increasing K.

An artificial model behaves similarly, except that the "direction of the zip" is different. The proof of the following theorem is analogous to that of *Theorem 1*.

Theorem 2

Assume that system (7) satisfies conditions (1)–(6) and (10)–(12) and is artificial. Then there exist $\lambda < K_1 < K_2 < \infty$ such that for $K \in (\lambda, K_1)$ all points of segment L_K are stable in the Ljapunov sense and L_K is an attractor of the system. For $K \in (K_2, \infty)$ the system has no stable equilibrium in the closed positive octant of S, x_1, x_2 space. For $K \in (K_1, K_2)$ the point $[\lambda, x_1(K), x_2(K)]$ divides L_K into two parts (one of which may be empty): the points of L_K to the right of this point are unstable, the points to the left are stable in the Ljapunov sense, and the part of L_K to the left of this point is an attractor of the system.

Assume that (19) holds, i.e., the motion of $[\lambda, x_1(K), x_2(K)]$ along the (moving) segment L_K is steady (from left to right in the natural case, and from right to left in the artificial case). Let c be an arbitrary nonnegative number or infinity. For $K > \lambda$ on the segment L_K there is exactly one point (λ, ξ_1, ξ_2) such that $\xi_2 / \xi_1 = c$. We say then that at level K the equilibrium (λ, ξ_1, ξ_2) realizes the ratio c of predators. If this equilibrium is stable in the Ljapunov sense {i.e., it is to the right of the point $[\lambda, x_1(K), x_2(K)]$ in the natural case and to the left of the point $[\lambda, x_1(K), x_2(K)]$ in the artificial case} then we say that at level K the ratio c of predators is stable (see *Figure 2*).

Corollary 3

If the conditions of *Theorem 1* or *2* and expression (19) hold and K_1 and K_2 have the meaning established in these theorems, then for $\lambda < K < K_1$ all ratios of predators are stable and for $K > K_2$ no ratio is stable. In the case where K is increased from K_1 to K_2 the ratios ξ_2 / ξ_1 are continuously destabilized so that in the natural case destabilization starts with high ratios and ends with low ratios and in the artificial case the opposite holds. In the natural case K_1 and K_2 are the unique roots of the equations $x_1(K) = 0$ and $x_2(K) = 0$ respectively. In the artificial case K_1 and K_2 are the unique roots of the equations $x_2(K) = 0$ and $x_1(K) = 0$ respectively.

The bifurcation, i.e., the loss of stability of the two end points of segment L_K, can be characterized. This could be done in the general case; however, the picture is clearer if (19) is assumed, and this we shall do. For $K > \lambda$ system (7) has an equilibrium in the interior of the positive quadrant of each coordinate plane S, x_i $(i = 1, 2)$, i.e., $[\lambda, \gamma\lambda g(\lambda, K) / p(\lambda, a_1), 0]$ and $[\lambda, 0, \gamma\lambda g(\lambda, K) / p(\lambda, a_2)]$. These coordinate planes are invariant manifolds of system (7), and the restriction of (7) to any of them is

$$\begin{aligned} \dot{S} &= \gamma S g(S, K) - x_i p(S, a_i) \\ \dot{x}_i &= x_i p(S, a_i) - d_i x_i \end{aligned} \tag{20}$$

The equilibrium of the two-dimensional system (20; $K \equiv K_i$) inside the positive quadrant of the S,x_i plane is $P_i(K) = [\lambda, \gamma\lambda g(\lambda,K)/p(\lambda,a_i)]$. It is easy to see either from the previous results or directly that $P_i(K)$ is asymptotically stable for $\lambda < K < K_{3-i}$ if the model is natural and for $\lambda < K < K_i$ if the model is artificial ($i = 1, 2$), where K_{3-i} and K_i are given in *Theorems 1* and *2* respectively. At $K = K_{3-i}$ (or $K = K_i$) a bifurcation characterized by the following theorem occurs (*cf.* Butler, 1983).

Theorem 4

If, in addition to (1)–(6), (10)–(12) and (19), we have $g, p \in C^4$, the equilibrium $P_i(K)$ of system (20) undergoes a Hopf bifurcation at $K = K_{3-i}$ in the natural case and at $K = K_i$ in the artificial case, where K_{3-i} and K_i are given in *Theorems 1* and *2* respectively (see also *Corollary 3*). The Hopf bifurcation is supercritical or subcritical depending on whether the expression

$$G(S) = \left[\frac{f''(S)}{p_i(S)p_i'(S)}\right]' p_i'^2(S)p_i'^2(S) + \left[\frac{f(S)p_i'(S)}{p_i''(S)}\right]' p_i''^2(S) \qquad (21)$$

is negative or positive, respectively, at $S = \lambda$ where $f(S) = Sg(S,K_{3-i})$ in the natural case, $f(S) = Sg(S,K_i)$ in the artificial case, and $p_i(S) = p(S,a_i)$.

Proof

The proof will be carried out for the natural case only. The artificial case is analogous; it is obtained by writing K_i for K_{3-i} everywhere. On linearizing system (20; $K = K_i$) at $P_i(K)$, the eigenvalues turn out to be

$$\mu_{1,2}(K) = \alpha(K) \pm i\beta(K)$$

where

$$\alpha(K) = -\frac{\gamma}{2p(\gamma,a_i)}\{-\lambda g'_S(\lambda,K)p(\lambda,a_i) + g(\lambda,K)[\lambda p'_S(\lambda,a_i) - p(\lambda,a_i)]\}$$

and

$$\beta(K) = \frac{1}{2p(\lambda,a_i)}\left[4\gamma\lambda g(\lambda,K)p^2(\lambda,a_i)p'_S(\lambda,a_i\right]$$
$$-\gamma^2\{-\lambda g'_S(\lambda,K)p(\lambda,a_i) + g(\lambda K)[\lambda p'_S(\lambda,a_i) - p(\lambda,a_i)]^2\}^{0.5}$$

It can be seen from equations (18), the proof of *Theorem 1*, and equation (19) that for $\lambda < K < K_{3-i}$ and $\alpha(K) < 0$, $\alpha(K_{3-i}) = 0$ and

$$\beta(K_{3-i}) = [\gamma\lambda g(\lambda,K_{3-i})p'_S(\lambda,a_i)]^{0.5}$$
$$=\{\gamma p(\lambda,a_i)[g(\lambda,K_{3-i}) + \lambda g'_S(\lambda,K_{3-i})]\}^{0.5} > 0 \tag{22}$$

A simple calculation yields

$$\alpha'(K_{3-i}) = \frac{1}{2}\gamma\lambda[g''_{SK}(\lambda,K_{3-i})$$
$$-g_S(\lambda,K_{3-i})[g'_K(\lambda,K_{3-i})][g(\lambda,K_{3-i})]^{-1}$$

Conditions (1), (3), and (19) imply $\alpha'(K_{3-i}) > 0$.

An application of the method given by Hassard *et al.* (1981) shows that the sign of the real part of the Floquet exponent of the bifurcating closed orbits is equal to the sign of $G(\lambda)$, and this proves *Theorem 4*.

Corollary 5

If $g, p \in C^4$ the conditions of *Theorem 1* (or *Theorem 2*) and inequality (19) hold and the expression $G(\lambda)$ in (21) is negative, then there exists a $\delta > 0$ such that, for $K \in (K_i, K_i + \delta)$, system (7) has a small-amplitude periodic solution around the unstable equilibrium $P_{3-i}(K)$ in the S, x_{3-i} plane [or around $P_i(K)$ in the S, x_i plane] which is locally unique in the corresponding coordinate plane and is orbitally asymptotically stable with respect to the restricted system $(20;K\equiv K_{3-i})$ or $(20;K\equiv K_i)$ $(i=1,2)$; its period is close to $2\pi/\beta(K_{3-i})$ [or $2\pi/\beta(K\,i)$].

Remark

If the model is ***degenerate*** then

$$\frac{p'_S(\lambda,a_1)}{p(\lambda,a_1)} = \frac{p'_S(\lambda,a_2)}{p(\lambda,a_2)} \tag{23}$$

It is easy to see that in this case the lines L_K and B_K exist and are parallel. For small $K > \lambda$ the line L_K is below B_K, and for large K the line L_K is above B_K. Thus, for small $K > \lambda$ all the points of the segment L_K are stable in the Ljapunov sense and L_K is an attractor of the system (7), and for large K no equilibrium is stable in the closed positive octant of S, x_1, x_2 space. The points of L_K change their stability behavior at identical values of K. If (19) holds then there exists a K_0 such that for $\lambda < K < K_0$ all points of L_K are stable, and for $K > K_0$ all points are unstable. *Theorem 4* and *Corollary 5* are also valid in the degenerate case with K_0 instead of K_i or K_{3-i}.

4. Examples

Let us assume that the prey's growth rate (without the factor γ) has the form

$$g(S,K) = 1 - (S/K)^u \qquad 0 < u \leq 1$$

which includes the logistic growth rate ($u = 1$). This function satisfies conditions (1)–(3); moreover, (19) is also valid. The following functional responses (see May, 1974) satisfy our conditions, while it can be shown that the first two make natural models and the third makes a degenerate model:

$$p(S,a) = \frac{A}{Ba + C}\,\frac{S}{S + a} \qquad \text{(Holling)}$$

$$p(S,a) = \frac{A}{Ba + C}[1 - \exp(-s/a)] \qquad \text{(Ivlev)}$$

$$p(S,a) = \frac{A}{Ba + C}S^q \qquad 0 < q < 1 \qquad \text{(Rosenzweig)}$$

where $A > 0$, $B \geq 0$, and $C \in \mathbf{R}$ can be determined so that $m_1 = A/(Ba_1 + C)$ and $m_2 = A/(Ba_2 + C)$ with arbitrarily prescribed values of $0 < m_1 \leq m_2$ and $a_1 > a_2 > 0$ (*cf.* (9) and (10)).

We were unable to find an artificial model satisfying conditions (4)–(6) in the literature. Nevertheless, such a model can be constructed as is shown by our final example:

$$p(S,a) = \frac{A}{Ba + C}\left[\ln(1 + S) + \frac{S}{1 + a}\right]$$

Acknowledgment

This paper was prepared while the author held a fellowship at the University of Alberta financed by the World University Service of Canada.

References

Butler, G.J. (1983), Competitive predator–prey systems and coexistence, in *Population Biology Proceedings, Edmonton 1982, Lecture Notes in Biomathematics 52*, pp 210–217 (Springer, Berlin).

Butler, G.J. and Waltman, P. (1981), Bifurcation from a limit cycle in a two predator–one prey ecosystem modeled on a chemostat, *J. Math. Biol.*, **12**, 295–310.

Farkas, M. (1984), Zip bifurcation in a competition model., *Nonlinear Anal. TMA*, **8**, 1295–1309.

Farkas, M. (1985), A zip bifurcation arising in population dynamics, in *10th Int. Conf. on Nonlinear Oscillations, Varna, 1984*, pp. 150–155 (Bulgarian Academy of Science, Sofia).

Hassard, B.D., Kazarinoff, N.D., and Wan, Y.–H. (1981), *Theory and Applications of Hopf Bifurcation* (Cambridge University Press, Cambridge, UK).

Hsu, S.H., Hubbell, S.P., and Waltman, P. (1978 a), Competing predators, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.*, **35**, 617–625.

Hsu, S.H., Hubbell, S.P., and Waltman, P. (1978 b), A contribution to the theory of competing predators, *Ecol. Monogr.*, **48**, 337–349.

Keener, J.P. (1983), Oscillatory coexistence in the chemostat: a codimension two unfolding, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.*, **43**, 1005–1018.

Koch, A.L. (1974 a), Coexistence resulting from an alteration of density dependent and density independent growth, *J. Theor. Biol.*, **44**, 373–386.

Koch, A.L. (1974 b), Competitive coexistence of two predators utilizing the same prey under constant environmental conditions, *J. Theor. Biol.*, **44**, 387–395.

May, R.M. (1974), *Stability and Complexity in Model Ecosystems*, p 82 (Princeton University Press, Princeton, NJ).

May, R.M. (1981). *Theoretical Ecology*, 2nd edn., p 79 (Sinauer, Sunderland, MA).

Smith, H.L. (1982), The interaction of steady state and Hopf bifurcations in a two-predator–one-prey competition model, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.*, **42**, 27-43.

Wilken, D.R. (1982), Some remarks on a competing predators problem, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.*, **42**, 895–902.

Chaos and the Theory of Elections

D.G. Saari
Northwestern University, Evanston, Illinois 60201, USA

1. Introduction

The concept of "chaos" is a fairly recent development, but already it plays an important, integral role in the study of dynamical systems. Using these mathematical techniques, we now understand how and why deterministic systems admit what appears to be highly random behavior (see, e.g., Devaney, 1986). For instance, these new approaches have enriched our understanding of numerical analyses and of physical and biological systems. Recently, as one might expect, chaos has been discovered in dynamical models developed for the social sciences (Saari, 1985a).

That chaos has and will increasingly be found in dynamical models in a wide variety of research areas is not surprising. What may be unexpected is that certain problems previously viewed only in a "static" context may also be analyzed in this way. By embedding these static problems in a quasi-dynamical framework, it is possible to understand, to relate, and to extend paradoxes that occur in decision theory, probability, integer programming, and several other areas. In this way, many new and unexpected paradoxes are exposed. Moreover, by using the development of dynamical systems as a guide, a new program emerges for the analysis of these topics. The purpose of this paper is to illustrate this with a particular example. I will describe the recent resolution of a long-standing problem from decision theory concerning elections.

A type of difficulty encountered with elections is illustrated by the following. Suppose a committee of nine voters is to choose among three candidates A, B, and C. Suppose that four rank the candidates as $A > C > B$, three rank them as $B > C > A$, and the last two rank them as $C > B > A$. By using the standard plurality election system, where each voter votes for his or her first-place candidate, the election result is $A > B > C$ with the tally of 4:3:2. However, this ranking is in conflict with the fact that a majority of these voters (five of the nine) prefer the second-ranked candidate B to the top-ranked candidate A. A more striking inconsistency is that two thirds of the voters prefer the bottom-ranked candidate C to B, and a majority prefers C to A. These transitive, binary rankings define a ranking, $C > B > A$, that is the exact reversal of the election ranking. More seriously, C is the majority candidate (who wins a majority vote in any pairwise comparison), but is ranked last in the election. The antimajority candidate A (who loses in any pairwise election) is ranked top.

During the American and French revolutions, procedures to aggregate individual preferences into a group ranking were advanced on both sides of the Atlantic Ocean. Partially stimulated by this discussion, examples of the above type were discovered. It was quickly recognized that these paradoxes occur because each voter can register only his or her top-ranked candidate. In his influential work, Borda (1781) proposed the simple solution of having each voter specify his or her ranking of the N candidates. Then, in the tallying process, $(N - j + 1)$ points are assigned to a voter's j-ranked candidate, $j = 1, \ldots, N$. A candidate's final ranking is determined by the number of points cast for him or her. (Actually, Borda's proposal was for $N = 3$.) For instance, in the above example, the Borda ranking of $C > B = A$ with a tally of 20:17:17 is closer to the apparently correct one.

This seems to be a reasonable solution, but it is not free from defects. Criticism of Borda's method, raised by Laplace and others, centered on three principal points, two of which I discuss here. The first concerns the weights used in the tallying process; what is the justification for adopting this particular choice? Instead of using (3,2,1), why not use (4,2,1)? The weights that could be used must satisfy certain obvious conditions. Any choice can be viewed as defining a *voting vector* $\mathbf{W} = (w_1, \ldots, w_N)$ where we impose a monotonicity assumption, $w_j \geq w_{j+1}$, $j = 1, \ldots, N - 1$. In order for the tally to distinguish between candidates, we require $w_1 > w_N$. For convenience, and because it always is satisfied in practice, we also assume that each w_j is a rational number. In the tally, w_j points are assigned to a voter's j-ranked candidate, which defines a *weighted voting system*. [For instance, the usual plurality voting method is defined by the vector $(1,0, \ldots, 0)$, while Borda's method is defined by $(N, N - 1, \ldots, 1)$.] The question raised by Borda's critics can be viewed as seeking a criterion to justify the choice of a particular $\mathbf{W}$ from the infinite number of possibilities.

The second criticism was that Borda's method need not always elect a "majority" candidate. To see this, consider the example of 11 voters where six have the ranking $B > A > C$, four have the ranking $A > C > B$, while the last voter has the ranking $C > B > A$. The Borda election ranking is $A > B > C$ with the tally of 25:24:17. Yet, the second-ranked candidate B wins by a majority vote when compared with either A or C. Although B is a "majority" candidate, B is not top-ranked in the Borda count. (However, here the plurality ranking of $B > A > C$ agrees with the binary rankings.)

We now know much more about voting systems and the theory of elections. One of the most important results was found by Arrow (1963). He proved that there is not a procedure that satisfies certain simple, reasonable axioms. It is an immediate consequence of his result that the above examples are not isolated phenomena; for *any* choice of a voting vector, the group's ranking of some pair of candidates need not be consistent with how this same group ranks the pair within a set of three or more candidates.

While this and several other results have shed light on the points raised by critics of Borda, the basic issues remained unanswered until recently. The solution, which is described in part here, is accomplished by borrowing ideas from chaos and symbolic dynamics.

2. Symbolic Dynamics

Our analysis of voting is strongly influenced by ideas from the dynamics of iterative processes, such as the system

$$x_{j+1} = f(x_j) \qquad j = 0,1,2,\ldots \tag{1}$$

If $p = x_0$ is an initial condition, then the goal is to understand the trajectory $\{x_k\}$, i.e., the sequence

$$\{f^{(k)}(p)\} \qquad k = 0,1,\ldots \tag{2}$$

where $f^{(k)}$ is the k-fold composition of f with itself. An important device used to describe chaos is *symbolic dynamics*. In this procedure, certain regions of state space are partitioned. Each partition set is labeled with a unique symbol, e.g., a letter from the alphabet. Then, for a specified initial condition, the trajectory of equation (2) defines a sequence of symbols where the $(k+1)$th entry is the symbol of the partition set containing $f^{(k)}(p)$. Such a sequence is called a "word" and the set of all possible words is called a "dictionary". By knowing which words are in a dictionary, we obtain a deeper understanding of the dynamics of the system. The extreme case is where all possible words occur; this is a form of chaos.

We use this approach of symbolic dynamics to address the critics of Borda and to analyze voting systems. To do this, a "dynamical" program is designed for elections. The procedure is determined by a set of functions $\{f_k\}$ that specify the election results. The "trajectories of voting" will assume the form of

$$\{f_k(p)\}, \qquad k = 1,\ldots,T \tag{3}$$

where the index identifies the subset of candidates being ranked.

The purpose of voting is to determine the group's rankings for specified subsets of candidates. From the set of N candidates $\{a_1,\ldots,a_N\}$, there are $2^N - (N+1)$ subsets with two or more candidates. Let F be a family of some of these subsets; that is, let $F = \{S_1,\ldots,S_T\}$ where S_j is a subset of two or more of the candidates and where $S_j \neq S_k$ if $j \neq k$. For instance, a family is used in the above examples; this family consists of all $N(N-1)/2$ pairs of candidates and the set of all N candidates. The criticisms of the plurality vote and of Borda's method are based on showing that the election results over the various subsets of this family are inconsistent. Other criticisms can be based on demonstrating inconsistent election results over other subsets of candidates; the choice of the subsets defines the family F.

Let a family $F = \{S_1,S_2,\ldots,S_T\}$ be given. Toward the goal of determining all of the election results, let R_j be the set of all possible linear rankings

of the candidates in S_j, $j = 1, \ldots, T$. [For instance, if $S_j = \{a_1, a_2\}$, then $R_j = \{a_1 > a_2, a_1 = a_2, a_1 < a_2\}$.] These rankings play the role of "symbols". Let $\mathbf{W}_F = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_T)$ be a set of T voting vectors where $\mathbf{W}_j$ is used to tally the ballots for the candidates in S_j, $j = 1, \ldots, T$. These vectors, which define the election procedure for each subset of F, describe the "dynamics", i.e., the mappings of equation (3).

To complete the description, recall that a "profile" of voters p is a listing for each voter of his or her ranking of the candidates. If P denotes the space of all possible profiles, then the election result for S_k is given by

$$f_{kW} : P \to R_k \tag{4}$$

where f_{kW} is the obvious mapping defined by the tallying process with $\mathbf{W}_k$. This mapping is defined in the following way. Once $\mathbf{W}_k$ is specified and a profile p is given, the election tally can be computed. In place of the actual tally, we use a "symbol" – the ordinal election ranking of the candidates. This is $f_{kW}(p)$. Thus, for a specified $\mathbf{W}_F$ and for a given p, the election results over all subsets of F are given by

$$\{f_{kW}(p)\}, \qquad k = 1, \ldots, T \tag{5}$$

To illustrate this, consider the example where $N = 4$, $F = \{\{a_1, a_2\}, \{a_2, a_3\}, \{a_1, a_2, a_4\}, \{a_1, a_2, a_3, a_4\}\}$, and $\mathbf{W}_F = (1,0;1,0;3,2,1;1,0,0,0)$. This choice of $\mathbf{W}_F$ means that the first two subsets are ranked by majority vote, the third subset is ranked by Borda's method, and the last subset is ranked by a plurality vote. Suppose the profile p is given by five voters with the ranking $a_1 > a_4 > a_2 > a_3$, four voters with the ranking $a_2 > a_3 > a_4 > a_1$, and four voters with the ranking $a_3 > a_4 > a_2 > a_1$. Then, $f_{jW}(p)$ is the election ranking of S_j in F, $j = 1,2,3,4$, when the ballots are tallied with the specified voting vector. The outcome for p is $f_{1W}(p) = a_2 > a_1$ with a tally of 8:5, $f_{2W}(p) = a_2 > a_3$ with a tally of 9:4, $f_{3W}(p) = a_4 > a_2 > a_1$ with a tally of 30:25:23, and $f_{4W} = a_1 > a_2 = a_3 > a_4$ with a tally of 5:4:4:0. The sequence defined by equation (5) is

$$(a_2 > a_1,\ a_2 > a_3,\ a_4 > a_2 > a_1,\ a_1 > a_2 = a_3 > a_4) \tag{6}$$

For a given family F, let D_F be the Cartesian product $R_1 \times \cdots \times R_T$. D_F is a universal set; an element of D_F is a sequence of rankings, one for each of the T sets in F, and all possible sequences are in D_F. As such, D_F contains all possible and impossible election results for the family F. For instance, if $F = \{\{a_1, a_2\}, \{a_2, a_3\}\}$, then D_F contains nine sequences ($|D_F| = 9$). If $N = 3$ and F is the set of all $2^3 - 4$ sets, then $|D_F| = 351$; if $N = 4$ and F is the family of all $2^4 - 5$ sets, then $|D_F| = 1\,686\,498\,489$.

To analyze the effects of a voting vector $\mathbf{W}_F$, we need to know all the possible election results for a family F. This is the subset of D_F consisting of all possible election results obtained by using $\mathbf{W}_F$:

$$D_{FW} = \{[f_{kW}(p)]\ k = 1, \ldots, T\ p \text{ is in } P\} \tag{7}$$

Following the lead of dynamical systems, we call D_{FW} the *dictionary* defined by $\mathbf{W}_F$ and a sequence in D_{FW} a *word*. For instance, the sequence given in equation (6) is one of the words in D_{FW} for the specified choice of $\mathbf{W}_F$.

To state our results, we need to describe an equivalence relation for $\mathbf{W}_F$. Our first statement weakly parallels ideas from "topological equivalence". We are interested in knowing when $\mathbf{W}_F$ and $\mathbf{W}_F''$ give rise to equivalent "trajectories"; that is, when do two sets of voting vectors always define the same election outcome for all choices of p? When they do, we call them *equivalent*. The proof of the following is a simple exercise.

Proposition 1

Suppose a family F is given. Two vectors, $\mathbf{W}_F = (\mathbf{W}_1, \ldots, \mathbf{W}_T)$ and $\mathbf{W}_F' = \mathbf{W}_1, \ldots, \mathbf{W}_T)$ are equivalent if and only if for each $j = 1, \ldots, T$, the three vectors $\mathbf{W}_j$, $\mathbf{W}_j'$, and $\mathbf{E}_s$ are all in the same two-dimensional subspace of an Euclidean space of the appropriate dimension. Here, the vector dimension of $\mathbf{E}_s$ agrees with $|S_j|$, the cardinality of S_j.

Because of this equivalence, we assume that if S_j has only two candidates, then the assigned voting vector is (1,0). This vector determines the tally for a majority vote. Also, it is easy to show that if the difference between successive weights of a voting vector is the same constant, then this vector is equivalent to a Borda vector, e.g., (45,20,−5) is equivalent to (3,2,1). If all of the vector components of $\mathbf{W}_F$ are either (1,0) or equivalent to a Borda vector, denote it by $\mathbf{B}_F$. Denote the dictionary for $\mathbf{B}_F$ as $D_{F\mathbf{B}}$.

3. Election Results

One way to address the questions raised by Borda's critics is to compare the dictionaries D_{FW} and $D_{F\mathbf{B}}$. For instance, suppose one dictionary admits more words than the other. This means that the associated voting method admits more "inconsistent" election results for F than the other method. In other words, as in dynamical systems, the larger the dictionary, the more "chaotic" are the possible election outcomes. On the other hand, a method with a smaller dictionary is more predictable; it is also more consistent. Thus, for F, we analyze the relationship of $D_{F\mathbf{B}}$, D_{FW}, and D_F.

It is easy to construct examples of a family F where $D_F = D_{FW}$ for any choice of $\mathbf{W}_F$. This is true for $F = \{\{a_1,a_2,a_3\}, \{a_4,a_5,a_6\}\}$. The reason is that the two sets have no overlap, so there is nothing to compare. This motivates the following definition.

Definition 1

A family F satisfies the ***binary inclusion property*** (b.i.p.) if (i) F contains at least one set with three or more candidates and (ii) if S_j in F has three or more candidates, there is a pair of candidates in S_j that is in some other set in F.

A one-set family, the family consisting of all $N(N-1)/2$ pairs of candidates, and the above two-set family do not satisfy the b.i.p. On the other hand, the family consisting of all pairs of alternatives and $\{a_1,a_2,a_3\}$ does satisfy the b.i.p. Indeed, any subfamily of F that includes this set and $\{a_1,a_2\}$ satisfies the b.i.p. $F = \{\{a_1,a_2,a_3\}, \{a_4,a_5,a_6\}, \{a_4,a_6,a_8\}\}$ does not satisfy the b.i.p., but the subfamily consisting of the last two sets does. Our first result is a negative one.

Theorem 1 (Saari, 1984, 1985b)

Assume there are $N \geq 3$ candidates:

(1) If there is no subfamily of F that satisfies the b.i.p., then for any choice of $\mathbf{W}_F$, $D_F = D_{FW}$.
(2) Let F be the family of $N-1$ subsets $S_{j-1} = \{a_1, \ldots, a_j\}$, $j = 2, \ldots, N$. For any choice of $\mathbf{W}_F$,

$$D_F = D_{FW} = D_{F\mathbf{B}} \tag{8}$$

There are several immediate consequences of this theorem. The obvious one is that for families of the above type the election results can be as "chaotic" as desired. Because $D_F = D_{FW}$, *any* word can be realized by some profile of voters. Thus, there exist paradoxes that are far more complicated than previously suspected; anything is possible! For instance, this means that there exists a profile of voters with the election rankings $a_1 < a_2 < a_3$, $a_3 < a_4 < a_5$, and $a_5 < a_6 < a_1$. (This family does not satisfy the b.i.p.) It follows from the second statement that there exists a profile of voters such that when, for the $1 > \cdots > a_j$ if j is even, but it is $a_j > a_{j-1} > \cdots > a_1$ if j is odd. Examples of this type are not the fault of Borda's method; there is no possible way to select the voting vectors to avoid these negative conclusions. Any choice leads to the same conclusion.

Another consequence of the second statement pertains to the standard procedure of "runoff" elections. This is where the final result is based upon an iteration procedure where, at each step, the "last place" candidate is dropped and the remaining set of candidates is reevaluated. This procedure defines a nested family of sets, so the family described in the second part of the theorem is a natural one to consider. But, according to this theorem, there exist profiles of voters so that when each set S_{j-1}, $j = 2, \ldots, N$, is ranked, a_1 is always ranked second to last and a_j is ranked in last place. (Thus, a majority of these voters prefer a_1 to a_2.) According to the procedure, candidate a_1 is always (nearly) advanced to the next stage. Ultimately, she or he is the winner.

This result offers no useful distinction between the Borda count and any other weighted voting method. Distinctions do emerge for other families.

Theorem 2 (Saari, 1985b)

(1) For any family F and any $\mathbf{W}_F$, $D_{F\mathbf{B}}$ is a subset of $D_{F\mathbf{W}}$.
(2) Suppose $F = \{S_1, S_2, \ldots, S_T\}$ satisfies the b.i.p. If there is a $\mathbf{W}_F$ such that $D_{F\mathbf{B}}$ is a proper subset of $D_{F\mathbf{W}}$, then for *all* choices of $\mathbf{W}_F \neq \mathbf{B}_F$, $D_{F\mathbf{B}}$ is a *proper* subset of $D_{F\mathbf{W}}$. If $\sum (|S_j| - 1) > N(N-1)/2$, then F has this property. In particular, for $N > 3$, this is true for the family of all $2^N - (N + 1)$ subsets.
(3) Corresponding to a family F is a Euclidean space E_F; this is the space that contains all vectors $\mathbf{W}_F$. In E_F there is an open, dense subset C_F such that $D_F = D_{F\mathbf{W}}$ if and only if $\mathbf{W}_F$ is in C_F. In particular, if the voting vector components of $\mathbf{W}_F$ all correspond to a plurality vote, then $\mathbf{W}_F$ is in C_F.
(4) If $N > 3$ and if the family F consists of $\{a_1, \ldots, a_N\}$ and all pairs of candidates, then C_F contains all of the vectors except $\mathbf{B}_F$.

In the proof of this theorem, $\mathbf{W}_F$ is treated as a parameter. As the results suggest, a type of singularity emerges. As one might expect, the conclusions are similar, in flavor, to those from singularity theory. For instance, these singularities define a stratified structure for the set of voting vectors. This stratification is given by the structure of the boundary of C_F defined above. As the choice of $\mathbf{W}_F$ varies through this stratification, the dictionaries $D_{F\mathbf{W}}$ are nested.

This theorem answers the first criticism of Borda's method. If we wish the choice of weights used in the tallying process to minimize possible and apparent inconsistencies in the elections over a given family of subsets, then it follows from Sections 1 and 2 that the *unique* choice of a tallying method is the Borda count. The other methods define dictionaries with more words; hence they admit more inconsistent election rankings. Indeed, from Section 3, the generic situation is that *any imaginable, chaotic outcome can occur*. The currently used plurality voting system is in this setting, as illustrated by the following corollary.

Corollary 1

Let $N \geq 3$ and let F be the family of all $2^N - (N + 1)$ subsets. For each of these subsets, arbitrarily select a ranking of the alternatives. There exists a profile of voters such that their plurality ranking of each of these subsets is the selected ranking.

Other consequences pertain to runoff elections, etc.

Corollary 2

Let $N \geq 3$. There exist profiles of voters so that the majority winner is always plurality ranked in last place in each of the subsets of more than two

candidates. Thus, for any runoff election that eliminates the bottom-ranked candidate, the majority winner for this profile will lose. Also, there exists a profile of voters so that for each subset of three or more candidates, the antimajority candidate is top ranked.

The second criticism of Borda's method was based on the fact that it may fail to rank a majority winner first. It follows from the fourth part of *Theorem 2* that there is *no* weighted voting system that will do this. Indeed, for any other choice of $\mathbf{W}_F$, $D_{F\mathbf{W}} = D_F$. As a consequence, any chaotic outcome is possible. The following is just one type of such outcome.

Corollary 3

Let $N \geq 3$ and let s_1 and s_2 be different integers between 1 and N. For any choice of a weighted voting system that is not a Borda system, there is a profile of voters such that the majority winner is ranked s_1th and the antimajority candidate is ranked s_2th.

The extreme case, where $s_1 = N$ and $s_2 = 1$, is illustrated by the introductory example.

It is possible to find other criticisms of Borda's method. This might be done by finding a Borda election ranking for a certain family of subsets of candidates that is viewed as being inconsistent. It follows from the second statement of *Theorem 2* that for *any* choice of $\mathbf{W}_F$ there is a profile of voters that defines the *same* election ranking. This is because $D_{F\mathbf{B}}$ is a subset of $D_{F\mathbf{W}}$ for any choice of F and for any choice of $\mathbf{W}_F$. Thus, the Borda system is the "most consistent" of the weighted election processes.

Because there are problems with voting methods, it is natural to consider only how candidates fare when evaluated against each other in pairwise comparisons. This was the approach advocated by Condorcet, another eighteenth century French mathematician who has had an important influence on the theory of elections. For instance, he advocated that a majority candidate should be declared the winner (Condorcet, 1785). Another approach might be to use an *agenda* for a meeting. This is where the N candidates or alternatives are listed, say as $[a_1, a_2, a_3, \ldots, a_N]$, and then pairwise compared according to this listing. The first two candidates are first compared, then the majority winner is compared with the third listed alternative, etc. Any such "binary method" can lead to inconsistencies; this is illustrated with the following two corollaries. [These results are based on the family of all $N(N-1)/2$ candidates. This family has no subfamily satisfying the b.i.p.]

Corollary 4

Let $N \geq 3$ and let F be the family of all $N(N-1)/2$ subsets of candidates. The election procedure, given by $\mathbf{W}_F$, is a majority vote for each subset; then $D_{F\mathbf{W}} = D_F$. That is, all possible cycles, subcycles, etc., of the rankings of the candidates are possible through majority vote.

Corollary 5

(1) For $N \geq 3$, there exists a profile of voters such that $a_j > a_{j+1}$, but $a_N > a_1$. For the remaining pairs of alternatives, $a_j > a_{j+k}$ if and only if k is an odd positive integer, and $a_j < a_{j+k}$ if k is an even positive integer.
(2) For this profile of voters, there is no majority and no antimajority candidate.
(3) For this profile of voters and for each choice of j, a_j is the winning candidate for the agenda $[a_{j+1}, a_{j+2}, \ldots, A_j - 1, a_j]$.

Sharper conclusions about election results can be found, and the interested reader should consult Saari (1985b). For instance, it is possible to characterize the set $D_{F\mathbf{B}}$, from which results of the following type can be found (which should be compared with the above).

Theorem 3

The Borda count is the unique weighted voting method that never ranks a majority winner in last place and an antimajority candidate in first place.

It is possible to characterize the sets $D_{F\mathbf{B}}$ and $D_{F\mathbf{W}}$ for any choice of F and $\mathbf{W}_F$. In this way, all possible election results can be found.

As a last illustration, suppose $N = 6$ and F is the family of all $2^6 - 7$ subsets of two or more candidates. The second part of *Theorem 2* asserts that $D_{F\mathbf{B}}$ is a proper subset of $D_{F\mathbf{W}}$ for any choice of $\mathbf{W}_F$. But, suppose there are only a few more words in $D_{F\mathbf{W}}$ than in $D_{F\mathbf{B}}$; then the above results, which favor Borda's method, lose much of their force. After all, this means that there is only a small additional number of "inconsistent" election results over those obtained with a Borda method. This is not the case; $D_{F\mathbf{B}}$ can have a significantly smaller number of words than $D_{F\mathbf{W}}$ for any $\mathbf{W}_F \neq \mathbf{B}_F$. For instance, suppose that none of the subsets of F are ranked with Borda's method. Then, the *very best* choice of a $\mathbf{W}_F$ that can occur has *each* word in $D_{F\mathbf{B}}$ replaced with more than 10^{18} words in $D_{F\mathbf{W}}$!

Acknowledgments

This is a written version of a talk given in September, 1985, at a conference in Sopron, Hungary, sponsored by SDS of IIASA. This research was supported, in part, by NSF grant IST 8415348. Some of this work was done while I was visiting the SDS at IIASA in Austria during September 1984.

References

Arrow, K. (1963), *Social Choice and Individual Values*, Monograph 12 (Cowles Foundation for Research in Economics, New Haven, CT).

Borda, J.-C. de (1781), Mémoire sur les élections au scrutin, *Histoire de l'Academie Royale des Sciences* (Paris).

Condorcet, Marquis de (1785), Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix (Paris).

Devaney, R. (1986), *An Introduction to Chaotic Dynamical Systems* (Benjamin, Menlo Park, CA).

Saari, D. G. (1984), The ultimate of chaos resulting from weighted voting systems, *Advances in Applied Mathematics*, **5**, 286–304.

Saari, D. G. (1985a), Price dynamics, social choice, voting methods, probability, and chaos, in C. Aliprantis, O. Burinshaw, and N. Rothman (Eds), *Advances in Equilibrium Theory*, Lecture Notes in Economics and Mathematical Systems, **244** (Springer, Berlin, Heidelberg, New York, Tokyo).

Saari, D. G. (1985b), *The Optimal Ranking Method is the Borda Count*, Collaborative Paper CP-85-4 (International Institute for Applied Systems Analysis, Laxenburg, Austria).

Spike-Generating Dynamical Systems and Networks

E. Labos
Semmelweis Medical School, 1st Department of Anatomy, Budapest, Hungary

1. Introduction

The word "spike" is a term for nerve cell electrical discharges. The description of a spike may be detailed or be restricted to the existence of a neural happening. Thus Hodgkin and Huxley (1952) described spikes by a nonlinear PDE coupled with 3 ODEs. However, McCulloch and Pitts (1943) applied logical constants which might also represent spikes. The choice of model strongly influences the speed of simulations and the tractability of problems. This paper suggests that both approaches may contribute to our knowledge. Moreover, present tools permit the design of intermediary models that are amenable to analysis – since generated by short algorithms – and which at the same time display complicated behaviors (May, 1976). Autonomous piecewise linear maps may be universal in generating interspike patterns or strange "neural" discharges (Labos, 1981; Nogradi and Labos, 1981). Two such systems, called polynomial spike generator (PSG) and universal pattern generator (UPG) will be discussed besides formal neural nets (FNNs).

2. Long Cycles Displayed by FNNs

FNNs are special $B^n \to B^n$ functions, where B^n is the n-dimensional Boolean cube. FNNs are realized by a square matrix N and a threshold vector T of real entries so that, for any $b \in B^n$, $f(b) = u(bN - T)$ holds, where $u(r) = 1$ if and only if $r > 0$ and $u(r) = 0$ otherwise.

It is easy to design a $B^n \to B^n$ function whose functional digraph consists of a single cycle of length $L = 2^n$, since any single cycle permutation over B^n gives a solution. The same question with the restriction that the cycle should be generated by an FNN (i.e., by an $<N,T>$ pair with iterations) represents a difficult problem. On this point, even a general existence theorem is lacking. A cycle is regarded as long if it is displayed by n neurons and its length exceeds half of the possible maximum, i.e., 2^{n-1}. The case of the absolute maximum (2^n) seems to be the hardest. A necessary condition and solutions of a few special cases (n=1,2,3,4,5,6) have been published (Labos, 1980, 1984).

Theorem 1

A vectorial Boolean function which is an FNN and whose state-transition graph consists only of cycles (i.e., it is transient-free – TF ; *Figure 1*) must have self-dual threshold gates as components which should be pairwise in Hamming distance of 2^{n-1}.

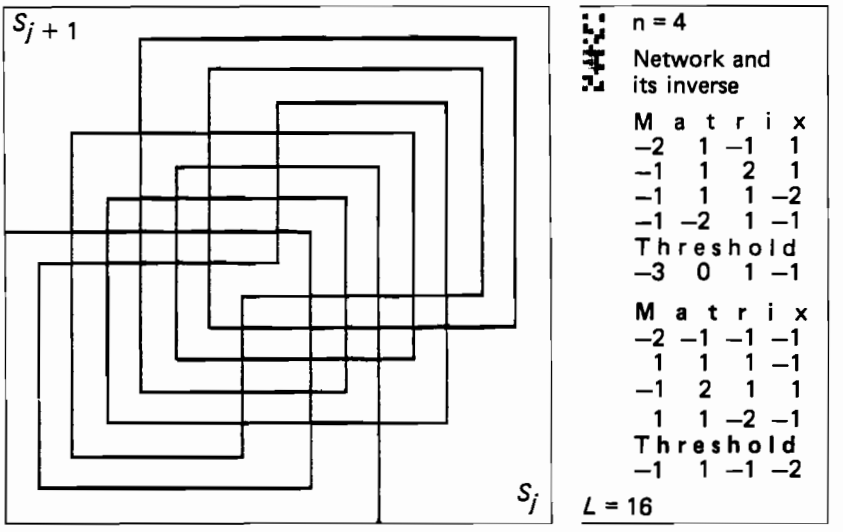


Figure 1 An FNN and its inverse. The same states flow in opposite directions along the trajectory.

2.1. Networks with one neuron: the case $n = 1$

All the four networks of one neuron are FNNs and two of them are invertible, self-dual, regular Hamming polyhedra. These are n-tuples of functions each with 2^{n-1} true vectors and pairwise in distance 2^{n-1}. The other two functions have no such properties.

2.2. The case $n = 2$

Of the 256 Boolean nets, 196 belong to the classes of FNNs of which 16 are self-dual: eight are invertible and eight have transient states. Of the 60 non-FNN nets, 16 are permutations on B^n. Forty-four items "have no face". Out of the 19 possible functional digraphs (Harary and Palmer, 1973), one cannot be found among FNNs: it is the permutation with (3,1) cycle decomposition. The cause of this lies in the self-dual property of TF FNNs.

2.3. The case $n = 3$. RC- and C-permutations

Since 104 of the 256 3-input truth functions are threshold gates (Muroga, 1971), it follows that of the 16 777 216 Boolean nets only 1 124 864 (= 104^3) are FNNs. Of these, 2744 are self-dual, of which 240 are TF (Labos, 1980). Of the 240 FNN permutations, only 48 have $L = 8$ (i.e., maximum length cycle) while 144 non-FNN self-dual and 39 936 non-FNN and non-self-dual TF Boolean nets exist. The self-dual permutations form a subgroup in the symmetric group of order 8!, since their product and inverse are also self-dual. In contrast, the 240 invertible FNNs do not form a subgroup; however, they generate the group of 384 self-dual permutations. Thus at $n = 3$ the majority of cases (15 606 024) lack the properties in question. The details of the classification applied here have been described by Labos (1984).

Theorem 2

The set of self-dual permutations on B^n represents a subgroup of the symmetric group of order $(2^n)!$.

Investigating the 22 partitions (see Andrews, 1976) of 8 in lexicographic order for the 240 relevant FNNs, the numbers in each partition are as follows: 48, 0, 8, 32, 0, 0, 0, 12, 0, 36, 0, 12, 32, 8, 0, 0, 0, 13, 28, 6, 4, 1. The sequence starts with (8), (7,1), ... and finishes with the (1,1,1,1,1,1,1,1) partition. Thus, for example, no (7,1) or (4,2,1,1) cycle decomposition is possible with FNNs of 3 inputs. A similar cardinality pattern holds for the 384 self-dual permutations, while the features prohibited here appear among the 39 936 remaining cases. Owing to this fact, "random net statistics" considered on Boolean nets (Gelfand, 1982) cannot be applied directly to the FNN subclass. The 48 optimal FNNs can be obtained from six "essential" cases simply by relabeling the variables. This operation corresponds to a simultaneous permutation of rows and columns in the matrix of the network. Lengths of state-transition cycles are invariant against these RC-permutations, while those of columns (C-permutations) strongly influence the cycle decomposition of states.

Although at $n = 3$ *Theorem 1* represents also a sufficient condition of reversibility, this does not hold from $n = 4$ onward.

2.4. The case $n = 4$. Connected and split numbers. Marginal states

The goal again is to synthesize and enumerate TF-FNN functions. This problem is still tractable: 104 self-dual FNN components may be listed; regular Hamming-polyhedral quartets may be selected. *Table 1* summarizes the results. The next problem is the synthesis of four FNN nets the behavior of which includes a long cycle ($L = 9,10,11,12,13,14,15$, or 16). Various methods

were applied as outlined below:

Table 1 List of primitive nets for generation of optimum cycle lengths by nets of four neurons.[a]

1 0 0 0	1 0 0 0	1 0 0 0	0–1 1 1	2–1–1–1	2–1 1 1
0 1 0 0	0 1 1–1	0 1 1–1	1 0 1–1	–1 2–1–1	–1 2 1 1
0 0 1 0	0 1–1 1	0 1 1 1	1 1 0 1	–1–1 2–1	1 1 2–1
0 0 0 1	0–1 1 1	0 1–1 1	–1 1 1 0	–1–1–1 2	1 1–1 2
0 0 0 0	0 0 0 0	0 1 0 0	0 0 1 0	–1–1–1–1	1 1 1 1
2–1–1–1	2–1–1 1	2 1 1 1	2 1 1–1	2–1 1–1	2–1–1–1
–1 2–1–1	1 2–1–1	–1 2–1–1	–1 2–1–1	1 2 1–1	–1 2–1–1
–1–1 2–1	1–1 2–1	–1–1 2–1	–1–1 2–1	–1–1 2–1	1 1 2–1
1 1 1 2	–1 1 1 2	–1–1–1 2	–1 1 1 2	1–1 1 2	1 1 1 2
0 0 0 3	–1 0 0 0	–1 0 0 0	–1 1 1–1	1–1 2–1	1 1 0–1
2–1–1–1	2–1 1 1	2–1–1 1	2–1 1–1	2 1–1–1	2 1 1–1
1 2 1 1	1 2–1 1	–1 2–1 1	1 2–1–1	1 2 1 1	1 2–1 1
1–1 2–1	1 1 2–1	–1–1 2 1	1 1 2 1	1–1 2 1	–1 1 2–1
1–1–1 2	–1 1 1 2	–1–1–1 2	1–1–1 2	–1 1–1 2	–1 1 1 2
2–1 0 0	1–1 1 1	–1–1–1 2	2 0 0 0	1 1 0 1	0 2 1 0
	2–1–1–1	2–1–1 1	2 1 1–1	2 1–1–1	
	1 2–1–1	1 2 1–1	–1 2–1 1	–1 2 1 1	
	1 1 2–1	1–1 2–1	–1 1 2–1	1–1 2 1	
	1 1 1 2	–1 1 1 2	1 1 1 2	1 1–1 2	
	2 1 0–1	1 0 1 0	0 2 1 0	1 1 0 1	

[a]The complete set of 22 pairs of matrices and thresholds, whose column or column/row permutations as well as partial negations lead to new transient-free network behaviors. Their total number is 233 × 384 of which about 10^5 have maximal ($L = 16$) cycle length. The corresponding numbers of Boolean nets are 16! or 15!.

(1) Independent subclocks with "local" cycle lengths which are primes to each other and whose total number of variables is equal to 4 yield solutions. For example, 2 FNNs, each with 2 neurons and with lengths $L = 3$ or $L = 4$ provide $L = 12$. This net splits into two disjoint parts. Lengths 10, 12, 14 are available in this way, assuming that certain lengths (the primitive or connected ones) were synthesized for $n = 1, 2, 3$.

(2) Primes (11,13) or prime powers (9,16) or a third category of numbers (here 15) require connected networks instead of split ones and a new method of synthesis.

Definition

The natural number L is called connected if $Q(L) < Q(z_1) + Q(z_2) + \cdots + Q(z_j)$, where $Q(x)$ is the upper integer of $\log_2(x)$ and the z_i are the relative prime factors of L in such a decomposition: e.g. $L = 45 = 9 \times 5$ is

connected since $Q(45) = 6 < Q(9) + Q(5) = 4 + 3 = 7$.

Remark

The diagnosis of connectedness of a number L needs $B(r) - 1$ investigations, where r is the number of prime divisors of L and $B(r)$ is the rth Bell-number. The smallest connected number is 15. In a broader sense, primes and prime powers will also be regarded as connected numbers. This definition is justified by the fact that such cycle lengths cannot be designed by a nonconnected FNN with a minimum number of variables (neurons). The nonconnected numbers will be called split numbers: e.g., even numbers between 2^{n-1} and $2^n - 1$ are split lengths.

The method designed to synthesize connected lengths is called the selective state-transition or shunt method and is based on special modifications of a start $< N,T >$ pair that generate lengths usually higher than the length to be constructed. Transitions of certain states are called marginal ones (*Figure 2*).

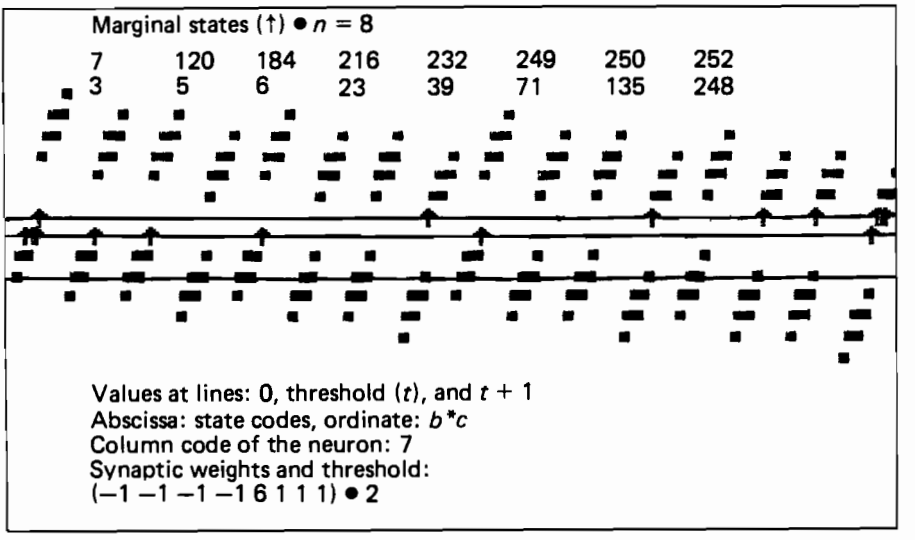


Figure 2 The distribution of "effects of states". Marginal states are the most sensitive to variations of network parameters. (Details in Section 2.4.)

Definition

Let $<N,T>$ be an FNN and $f = (c,t)$ one of its component threshold gates. The points $b \in B^n$ are called marginal with respect to f. The f are either true vectors, with effects equal to $U(f) = \min\{S * c\}$; or false vectors, with effects equal to $D(f) = \max\{K * c\}$; S and K are the sets of true or false vectors of f, and $*$ denotes an inner product. $U(f) - D(f) = g$ is the gap of f at the realization $f = (c,t)$.

We give an FNN for $n = 4$ which provides all cycle lengths if suitable modifications are applied: take the 17th matrix (in row 3) from *Table 1*; negate the 3rd and 4th functions; finally, apply the next column permutation – (4231). The state cycle is as follows: 0.10.9.14.13.7.4.12.15.5.6.1.2.8.11.3. By modifying 2 entries in certain columns and thresholds, all cycle lengths become available.

2.5. The Case $n = 5$. Codes of matrix columns, modifications, and states

Since this work deals with a large number of matrices of a special kind, it is important to introduce a short notation. Columns of certain square matrices are given by an array of decimal numbers: e.g., (11,20,19,8,10) is a 5×5 matrix code. How to decode? First, write $n-2$ as diagonal entries; here it is three. Secondly, write the binary form of the column code: e.g., the 1st column corresponds to 01011. Write '1's instead of nondiagonal '0's and leave '+1's unchanged. Compute now the value of the 1st threshold: it is the number of '1's in the column minus 1. Thus the first function is specified by (3,1,−1,1,1) and (2). Decoding all function codes, the partial negation has to be specified. The negation code 1 means 00001, a command to negate the 5th function. This is available by changing the signs of entries in the column and writing $-(t+1)$ instead of the original threshold t. The negated 5th function is given by (1,−1,1,−1,−3) and (−2). The next matrix transformation is a C-permutation. This is given by an array of column subscripts: e.g., (13542) means the (235) permutation with two fixed columns.

Now we deal with the M= (11,20,19,8,10) * neg(1) * (13542) specification. The result is an optimum-length-generating FNN with the following state sequence: (0.4.12.5.8.7.10.3.2.6.14.15.11.18.22.30.31.27.19.26.23.24.21.28. 29.25.17.16.20.13.9.1). These 32 numbers encode the 32 state-vectors if converted to binary form. Thus 01010 → 00011 is a state transition in this optimal cycle. On investigating the 120 column permutations and 32 negations, 3840 cases emerge, of which in this special case 352 have cycle length $L = 32$.

Our final goal is to give all those modifications of this matrix M by which specific shunts inside its cycle can be evoked. To do this, it is necessary to investigate the marginal states of each component as defined in Section 2.4. Each component has 5 upper and 5 lower marginal states: e.g., state 10 is upper marginal of the 5th function. If state 10 were jumped into state 2 instead of 3, then the cycle length would be $L = 31$. This is achieved by increasing the (5,5) entry and decreasing the (5,4) entry, together with increasing the threshold. The value of all changes is uniformly 0.5. This modification is coded as follows: $a = 5(5,4)$, which is a command to increase the 5th entry and decrease the 4th entry in the 5th column and also to increase the 5th threshold.

Now we list the modification codes for various connected lengths: $a = 5(5,4)$, $b = 4(3,1)$, $c = 1(5,2)$, $d = 3(1,3)$, $e = 1(4,5)$. These have to be applied to the negated–permuted matrix M as follows: $31 = Ma$, $29 = Mab$, $27 = Mc$, $25 = Mbc$, $23 = Mabcd$, $19 = Meb$, $17 = Mabde$. Thus to reach $L = 23$,

it is sufficient to modify 2–2 entries in the following columns: 5th (a), 4th (b), 1st (c), and 3rd (d). The realized shunts are: 10(3..)2, 19(26..)24, 14(15..)11, and 28(29..)13, respectively.

By these rather sophisticated procedures it has been essentially shown that all permitted cycle lengths may be implemented by an FNN of 5 neurons. The lengths not listed here are accessible by the independent sublock method using a disjoint collection of smaller networks.

2.6. The case $n = 6$. Structural stability of an FNN and its limits

The code of the start matrix is (31,15,7,3,1,0). That is, the value of all its diagonal entries is 4 while in the upper triangle there stand '– 1's and in the lower one '+1's. This matrix (defined for any n) will be referred to as matrix A. The 0-negation of the (123456) identical column permutation results in a behavior including one $L = 12$ cycle and 52 fixed points. Scanning through the 64 negations of this permutation, the length spectrum of cycles enclosing the state 0 contains 1,2,3,5,6,7,9,10,14,18,24. None of these are long enough. However, among the negations of other permutations – e.g., (2,6,5,1,3,4) – $L = 64$ appears 10 times besides 34,38,40,52, which are also long cycles.

Now we list certain C-permutations and negations of the A matrix which provide $L = 64$ cycles: (352614)*6, (365142)*54, (561342)*8, (542613)*4, (653142)*41, (562314)*23, (513264)*42, (615234)*34, (265134)*32, (564321)*5, (521364)*26, (654312)*24, (564321)*57, (546231)*46, (265134)*45, (635241)*58, (352614)*46, (251346)*7. The complete negations of any such nets also yield $L = 64$. Thus (251346)*56, the negation of the last item, also displays $L = 64$.

The next problem is the synthesis of the "strange" cycle lengths. The following procedure proved to be very effective: matrix $A \rightarrow$ neg(32) $\rightarrow$ (265134) column permutation $\rightarrow M$. Now 7 column modifications are listed: $a = 1(3,4)$, $b = 2(4,3)$, $c = 2(1,2)$, $d = 3(4,5)$, $e = 4(i,1)$, $f = 5(5,4)$, $h = 6(2,3)$. Applying suitable patterns of these to the network specified above, we obtain the connected cycle lengths as follows: 63 = Mb, 61 = Ma, 57 = Mc, 55= Mbf, 53 = Maf, 51 = Mch, 49 = Mcf, 47 = $Mafh$ or 47 = $Maef$, 45 = $Mceh$, 43 = Md, 41 = $Maefh$, 37 = Mdh. However, for $L = 59$ a new matrix was required: $A \rightarrow$ neg(5) $\rightarrow$ (564321) $\rightarrow N$. The modification for shunt is: $x = 5(3,4)$. The result is 59 = Nx. Experience shows that the modifiability of a given network is limited. The precise conditions of this phenomenon are not yet known. On the other hand, it is easy to prove that FNNs of fixed n are structurally stable in the following sense.

Theorem 3

An FNN is designed always with $<N, T>$ so that between the effects of upper and lower marginal states there lies a positive gap $g > 0$. All the matrix entries as well as the threshold may be modified by a quantity $\varepsilon > 0$ so that the behavior remains unchanged.

Remark

The term "effect" of state $b \in B^n$ is its inner product with column c. If a gap g is given, then $\varepsilon < g/(2n+2)$ must hold. Thus this kind of stability decreases if the number of neurons increases. The question is open as to how the ratio of the gap g and the range of effects could be maximized. The range is the difference of maximal and minimal effects when b runs over B^n. This problem is related to the maximization of structural stability. The practical consequences are evident: both the reliability and the tolerance against noise and vulnerability of formal neurons or technical threshold gates strongly depend on this issue. Related questions have been discussed by Muroga (1971) and others.

2.7. Networks with 7 formal neurons

As n grows, *Theorem 1* becomes less and less efficient. Nevertheless, a program (AUTODESIGN including NEXPER of Nijenhuis and Wilf, 1981) after some thousands of trials resulted in two solutions for $L = 128$. These are as follows:

(1) matrix B: (62, 30, 14, 6, 3, 1, 124), neg(1), (7365124).
(2) matrix C: (31, 64, 96, 48, 56, 60, 62), neg(1), (5627431).

Ultimately the matrix D – (63, 31, 15, 7, 3, 1, 0), neg(79), (5762314) – is of interest, and produces $L = 126$. The shunt method was applied to reach connected cycle lengths. A pattern of solutions is presented in detail below. Fifteen column transformations were necessary: $a = 7(5,6)$, $b = 7 = (1,2)$, $c = 6(3,4)$, $d = 5(5,4)$, $e = 4(6,7)$, $f = 4(2,1)$, $g = 5(5,6)$, $m = 2(2,3)$, $n = 1(2,3)$, $t = 3(4,5)$, $o = 4(1,2)$, $z = 7(7,4)$, $u = 5(6,7)$, $x = 6(5,2)$, $w = 1(1,2)$. The list of their applications to the B, C, and D networks is as follows: 127 = *Cu*, 125 = *Ba*, 123 = *Dx*, 121 = *Cuw*, 119 = *Bb*, 117 = *Bac*, 115 = *Bad*, 113 = *Be*, 111 = *Bbc*, 109 = *Bbd*, 107 = *Bf*, 103 = *Bde*, 101 = *Bga*, 99 = *Bcf*, 97 = *df*, 95 = *Bgb*, 89 = *Bge*, 85 = *Bmoz*, 83 = *Bgf*, 81 = *Becz*, 79 = *Bmnz*, 73 = *Bmez*, 71 = *Bgnz*, 67 = *fgt*.

2.8. The case of 8 threshold gates

Up to now the highest n at which the $L = 2^n$ optimum has been reached is 8. Here 5 solutions are presented:

(1) + (2) Start matrix, (127,63,31,15,7,3,1,0), i.e., the matrix A; C-permutation, (42856317), with 6 and 101 negations.
(3) The same start matrix with C-permutation (6834125) and 21 partial negation code.

(4) + (5) Start matrix, (63,191,31,15,7,3,1,0); C-permutation, (14856237); partial negation, 6 or 90.

A state-transition diagram is depicted in *Figure 3*, where the decimal code of the next state is plotted against the argument state code. Lines from (x,x) to (x,y) and from (x,y) to (y,y) have been drawn.

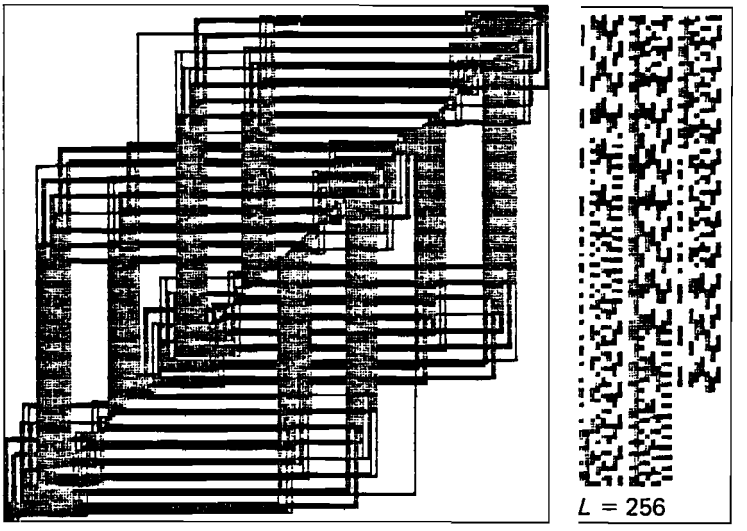


Figure 3 Optimal network behavior: $n = 8$. The decimal code of the next state is plotted against its predecessor. The punched tape of the states is shown on the right-hand side of the figure.

2.9. Large networks. Tractability

If an FNN of n neurons displays a cycle of length L – the case (n,L) – then a combination of two procedures permits construction of network sequences whose cycles increase rapidly. First, the enlargement of a net by state recognizer neurons permits an arbitrary jump from a state x to a state y if x is not an endpoint of the behavior graph. With a second enlargement, any length may be duplicated. This provides an $(n+2,2L)$ increase. A more rapid evolution is obtained by applying a recognizer to reach, from (n,L), the case $(n+1,L+1)$ and by using these two sublocks to reach $[2n+1, L*(L+1)]$. Iterations of these constructions provide net sequences. These procedures result neither in optimal nets nor in long cycles. However, the complete set of solutions up to $n = 8$ gives long L solutions up to $n = 70$, using sublocks of strange lengths. Moreover, the sublock method results in nondense wiring of nets with locally strong or even complete connections.

Numerous questions of optimal cycle design remain open. The most important is to find an iterative design which still keeps the cycles of the network sequence long or even maximal. The group structure involved in sets

of FNNs also represents a difficult subject (Biggs and White, 1979; Cameron, 1983; King, 1980). Enumeration (Goulden and Jackson, 1983; Harary and Palmer, 1973; Moon, 1970) of special FNNs is hindered by the fact that even the threshold gates have not yet been counted (Muroga, 1971). It would also be interesting to clarify the tractability status of threshold logic including a search of NP-complete problems (Garey and Johnson, 1979: see L09 in their catalog).

2.10. Taxons of boolean and threshold gates and nets

In this section some partially solved classification and enumeration problems are touched upon.

2.10.1.

Boolean sequences of fixed length can be classified according to their internal cycle length, recursive order, and number of '1's inside. These problems can be effectively handled with Möbius inversion and by a theorem of De Bruijn.

2.10.2.

Concerning $B^n \to B^n$ functions, the threshold gate property and self-duality as well as Chow-parameters seem to be important. Besides self-duality, the concept of antidual functions yields interesting results. A truth function is antidual if its dual pair is equal to its negation. By applying dual comparabilities, an exhaustive classification is available with 12 taxons. If at fixed n the sets of threshold gates (T), self-dual (D), and antidual (A) functions, and the two sets (V and W) of dual-comparable functions, or finally the set of functions with 2^{n-1} true vectors (H), are represented by logical variables, then the nonvoid and empty classes can be characterized by a "taxon function" as follows:

$$F(T,H,V,W,D,A) = \bar{T}.\bar{V}.\bar{W}.\bar{D} + \bar{H}.\bar{D}.(V \circ W).(T.A + \bar{A}) + H.V.W.D.\bar{A}$$

where "." = and, "+" = or, and " ∘ " = exclusive or. This expression consists of 12 terms determining 12 classes. For example, at $n = 4$ the sizes of these classes are: 104, 88, 88, 1, 1, 152, 70, 12 544, 5416, 5416, 184, 39 872. The last class is the largest in which none of the listed properties is satisfied (having no face).

2.10.3.

A taxonomy of similar principle was introduced for vector–vector Boolean functions by Labos (1984).

2.10.4.

Although both the census of finite automata and the census of nonlabeled loop-free functional digraphs (called functions – FDGs) have been completed (see Harary and Palmer, 1973), a large number of census problems related to the state-transition graphs of Boolean nets or FNNs remain open. A relatively easy one is the classification of FDGs according to the number of endpoints or initial states (i), true transients (t), nonfixed recurrent states (c), and fixed points (f) as parameters. The labeled case of this problem leads to a special counting of forests (of transients). If $x = i + t + c + f$ is the number of arguments, $m = c + f$ is the attractor's size, and $s = i + t$, then altogether $(1/6) * (x^3 + 5 * x)$ classes exist. Let us denote by $h(i,t,c,f)$ the size of such a class. Then

$$h(i,t,c,f,) = \sum_{k=1}^{i} F(s,k,i).m^k.\begin{bmatrix} x \\ m \end{bmatrix}\begin{bmatrix} m \\ f \end{bmatrix} P(c)$$

where $F(s,k,i)$ is the number of forests with s points, k denotes rooted tree components, and i the total number of endpoints in the forest. $P(c)$ is the subfactorial of c.

2.10.5.

A different classification problem emerges concerning the structure of FNN matrices and the corresponding wiring graph. The behavior is largely determined by the excitatory and inhibitory interconnections (i.e., by the + and – matrix entries) as well as by the pacemaker and nonpacemaker neurons. A neuron is a pacemaker if its threshold is negative. Thus each wiring structure corresponds to a digraph with 2-colored (labeled) points and 2-colored edges. Thus a wiring is defined here as a $2p2q$ labeled digraph. At $n = 4$, 218 digraphs exist. The numbers of $2p$, $2q$, or $2p2q$ labeled digraphs are approximately 3000, 20 000, or 360 000, respectively.

3. Spikes and Preturbulent Oscillations of a Polynomial Iteration

X, Y, Z, and W denote 4 successive values of a 3-order nonlinear iteration. The value W is computed from the previous ones as follows:

$$W = Z + a(Z - p)(Z - r)(Z - s) + bX^3 + cY^4 + k/(Z - s)$$

where $a = -1.2$, $p = 1$, $r = 0$, and $s = -0.2$ are fixed. At "standard" values $-b = -0.675$, $c = 0.225$, $k = 0.0016$ – spike oscillations appear whose otherwise not always constant "frequencies" may be controlled by k (from 0 to 0.06). If $b = -0.45$ and c runs from -0.5 to 0.35, the various new motions

emerge, the first return diagrams of which are depicted in *Figure 4*. The route to apparent chaos includes "toroidal" motion.

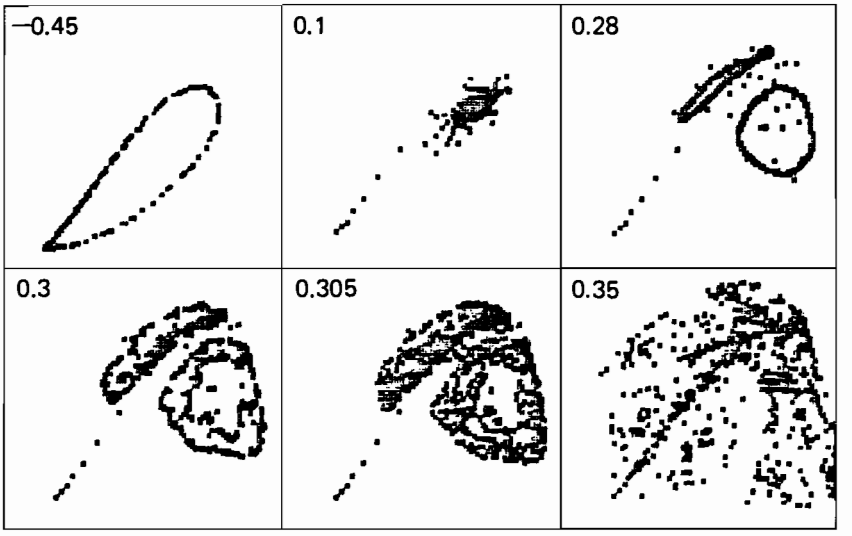


Figure 4 The value of the next state is plotted against its predecessor: 1000 iterations; the system of Section 3. The parameter c varies while $b = -0.45$, $k = 0.0019$. Magnifications for $c = 0.305$ and $c = 0.35$ are depicted in *Figure 7* after 10^4, 3×10^4, or 10^5 iterations.

A global view of the attractors of this system is obtained by scanning through a range of b or c parameters while others are fixed. *Figures 5* and *6* include spike oscillations, damping, intermittence, stable point attractors, or more and more irregular behavior. The hyperfine structure of these diagrams studied in micro-frames shows high numbers of special bands, and bifurcations at least up to 8 branches (see *Figures 5 – 7*).

The system was designed to simulate both normal and abnormal nerve cell discharges in a simple way. The irregular modes do not fit well to the known abnormal paroxysmal oscillations of real cells (*cf.* Holden, 1984), although the spike mode reproduces a wide spectrum of normal behavior.

4. Properties of a Universal Spike Pattern Generator (UPG) [Labos (1984)]

The system is a piecewise linear map on the interval [0,1]. In its standard version it is given by one parameter $a \in (0,1)$ and by two lines:

$$Y_{n+1} = mY_n \quad \text{if} \quad Y_n < a \qquad \text{and} \qquad Y_{n+1} = uY_n - u \quad \text{if} \quad Y_n \geq a$$

where $m = 1 + (1-a)^2$ and $u = 1/(1-a)$. Fixing $a = 0.1$, we have $m = 1.81$.

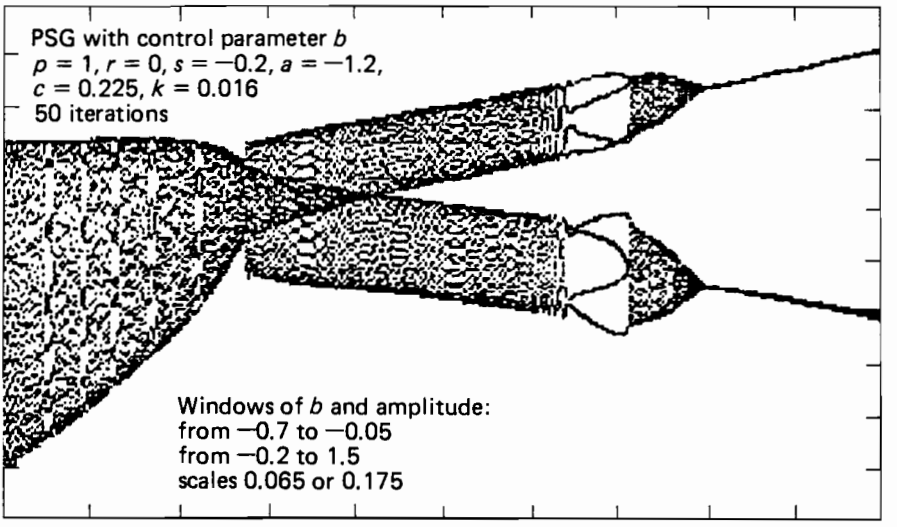


Figure 5 Variation of attractors of the PSG system when the parameter b is changed.

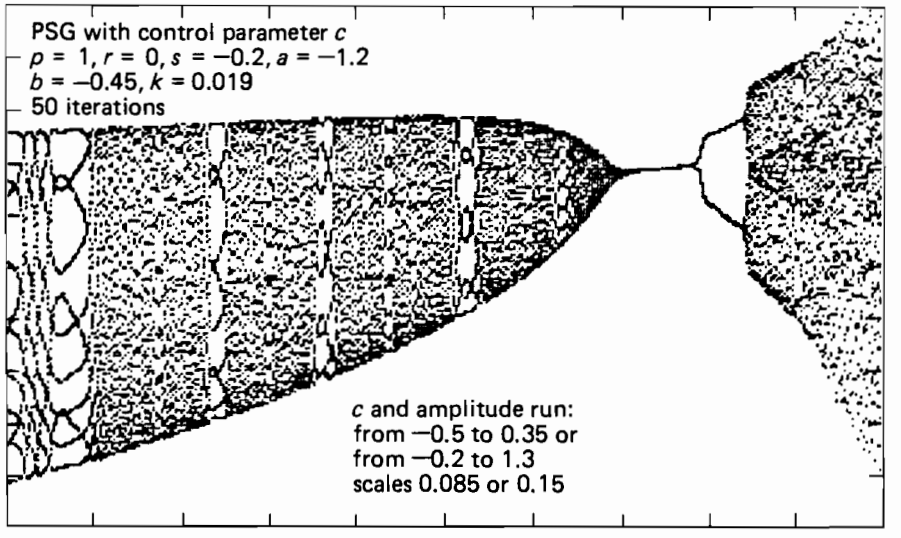


Figure 6 Variation of attractors of the PSG system when the parameter c varies. The number of iterations is altogether 16 000.

The two lines are denoted by L_1 and L_2.

In its extended version, m is independent of the threshold a (fixed at 0.1) and gives bounded motions if $10 > m > 0$. If $m > 1$, then this "neuron" is an oscillating pacemaker. For further purposes, its behavior was also defined below 0 by a third line: $Y_{n+1} = wY_n + e$, where the slope $w \in [0,1]$ and e is 0

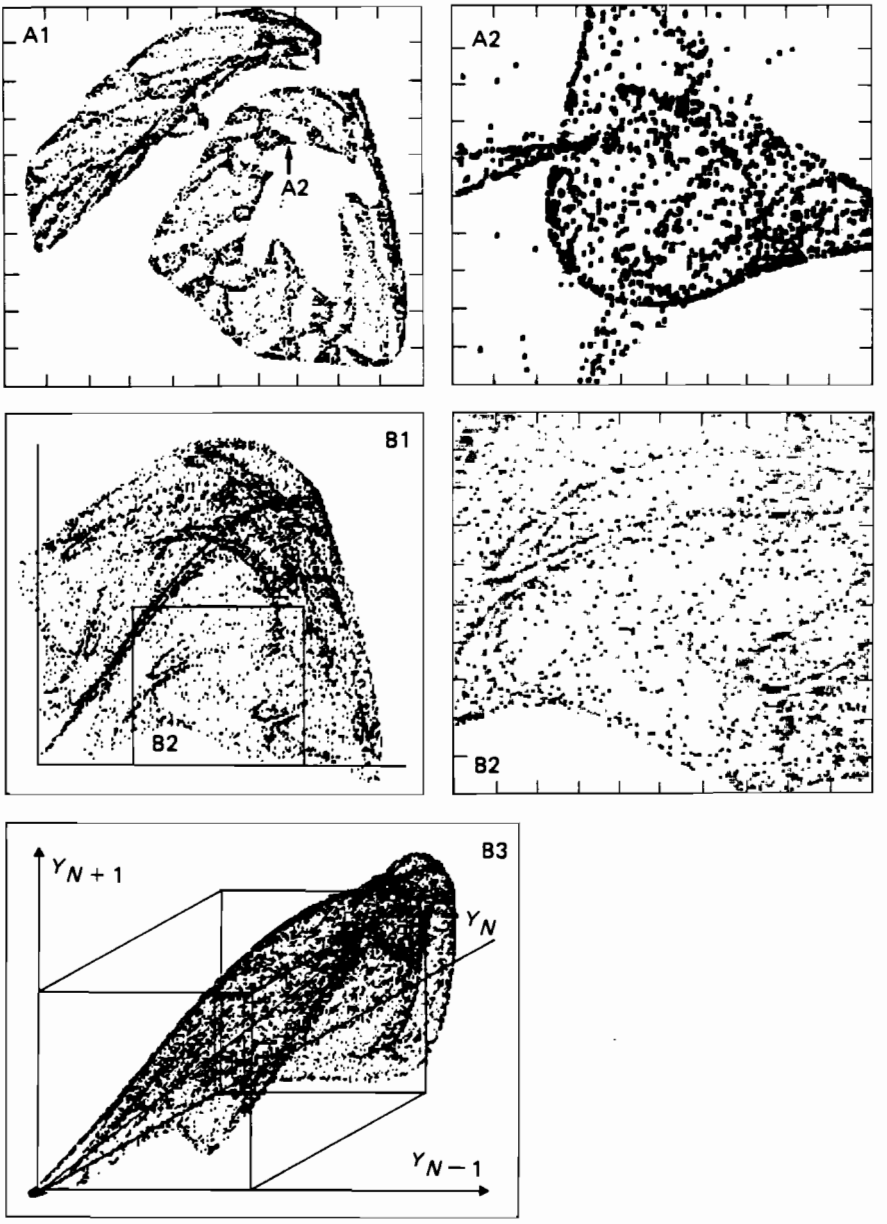


Figure 7 PSG system (Section 3): examples of strange attractors. Values of variables are plotted against their predecessors. The parameters a, p, s are fixed as written in the text; $k = 0.0019$, $c = -0.45$. At A and B the values of b are 0.305 or 0.35. Iterations: A1, 10 000; A2, 100 000; B1, 10 000, B2, 30 000; B3, 20 000.

or a small positive number. This makes possible the coupling of such units into networks.

A survey of some interesting properties follows:

(1) The iteration has two kinds of periodic solutions obtained by solving the following equations to initial value x: $x = xL_1^k L_2^2$. A periodic solution is called simple if between the start and return values only subthreshold

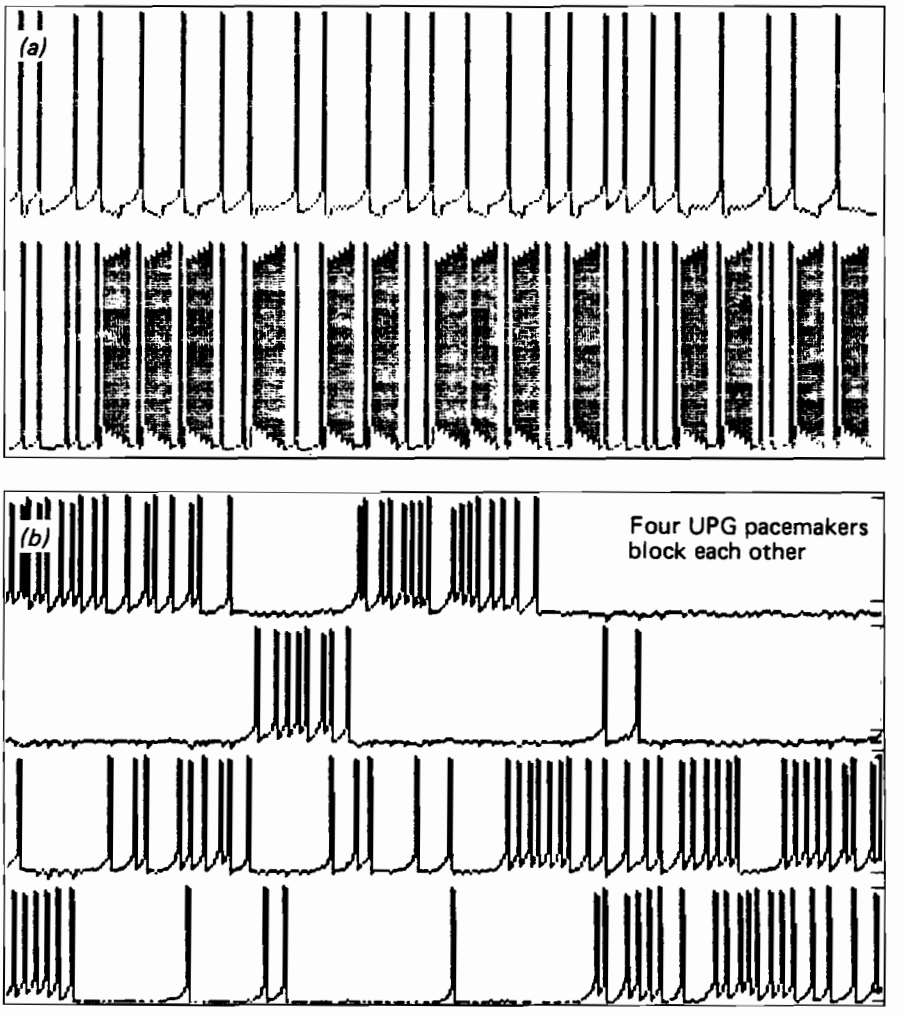


Figure 8 Coupling *UPG* modules into a network. The units are defined in Section 4. (*a*) The upper record is emitted by a pacemaker module ($m > 1$), which activates the lower bursting unit; the latter would be otherwise stable and silent; a negative inhibitory effect is fed back to the command neuron. (*b*) Four pacemaker UPGs are coupled by a completely inhibitory matrix; the resulting patterns of activity are depicted.

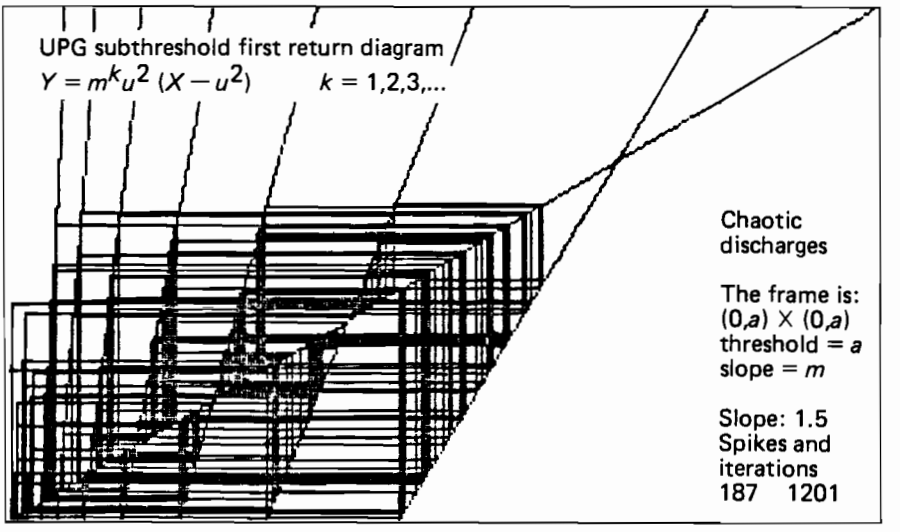


Figure 9 A special diagram whose fixed points correspond to the simple, and closed trajectories to the complex (i.e., spike-partitioned), periodic solutions of the UPG system (Section 4). Here an example of the third kind, chaotic motion, is displayed.

values occur. The solution is called complex if in a period the spikes generate a partition of iterations into a "pattern" of interspike intervals.

(2) The majority of solutions are aperiodic (starting with a nonrational initial value, the rational formula of periodic solutions inhibits periodicity).

(3) At reasonable values of m, periods of any natural number p can be obtained, except some smaller values. For a given p, usually more than one solution exists. The smallest interspike interval is 3 if $m = 1 + (1 - a)^2$.

(4) In simulations the periodic solutions show an easily computable finite lifetime. Trajectories started at different values diverge rapidly. Sometimes computer simulations display a persistent exact periodicity – an artifact with respect to the ideal case.

(5) Patterns of preassigned transients and prescribed spike-partition of the cycle may be designed.

(6) These units simulate surprisingly well real neural network properties if connected into small nets (*Figure 8*). The units have dead (or refractory) time, and may excite or inhibit each other according to the plan of the matrix. Their behavior remains bounded even after strong inhibition or excitation. The manner of coupling may be different: e.g., the influences are added to the autonomous values after (+ or −) amplifications or parameter controls are introduced to influence time-constants etc. The firing patterns in small networks can be predicted relatively well using the network

taxonomy described in Section 2.10.5. These taxons are valid also for nets of UPG:

(a) Connections are realized by + and − matrix entries.
(b) Pacemakers in FNNs have negative thresholds and in UPG nets $\underline{m} > 1$.

This latter corresponds to an unstable "resting" state. Since the piecewise linear units may be generalized to simulate a broader spectrum of normal and pathological excitable cell behavior (Nogradi and Labos, 1981; Labos, 1981), a flexible tool is available for rather quick computer simulations of either normal or abnormal network dynamics.

The UPG system can be regarded as chaotic (*Figure 9*) in the sense of Li and Yorke (1975; see also Misiurewicz, 1983). It even shows a special kind of self-similarity; however, an involvement of some Cantorian structure needs further analysis.

References

Andrews, G. E. (1976), *The Theory of Partitions* (Adison-Wesley, Reading, MA).

Biggs, N. L. and White, A. T. (1979), *Permutation Groups and Combinatorial Structures* (Cambridge University Press, Cambridge, UK).

Cameron, P. J. (1983), Automorphism groups of graphs, in L. W. Beineke and R.J. Wilson (Eds), *Selected Topics in Graph Theory* 2, pp 90–127 (Academic Press, New York).

Garey, M. R. and Johnson, D. S. (1979), *Computers and Intractability* (Freeman & Co, San Francisco, CA).

Gelfand, A. E. (1982), A behavioral summary for completely random nets, *Bull. Math. Biol.*, **44** (3), 309–320.

Goulden, I. P. and Jackson, D. M. (1983), *Combinatorial Enumeration* (Wiley–Interscience, New York).

Harary, F. and Palmer, E. M. (1973), *Graphical Enumeration* (Academic Press, New York).

Hodgkin, A. L. and Huxley, A. F. (1952), A quantitative description of the membrane current and its application to conduction and excitation in nerve, *J. Physiol.*, **117**, 500–544.

Holden, A. V. (1984), Why is the nervous system not as chaotic as it should be? *Abstract* from a Workshop held in Budapest.

King, R. B. (1980), Chemical applications of group theory and topology, *Theoret. Chim. Acta (Berl.)*, **56**, 269–296.

Labos, E. (1980), Optimal design of neuronal networks, in G. Szekely *et al.* (Eds), *Neural Communication and Control*, Adv. Physiol. Sci., 30, pp 127–153 (Pergamon Press, Oxford),

Labos, E. (1981), A model of dynamic behavior of neurons and networks, *Lecture Abstract of Annual Meeting of Hungarian Physiological Society, Budapest, I.S4*, p 117. (In Hungarian.)

Labos, E. (1984), Periodic and non-periodic motions in different classes of formal neuronal networks and chaotic spike generators, in R. Trappl (Ed), *Cybernetics and System Research*, 2, pp 237–243 (Elsevier, Amsterdam).

Li, T. Y. and Yorke, J. A. (1975), Period three implies chaos, *Am. Math. Monthly*, **82**, 985–992.

May, R. M. (1976), Simple mathematical models with complicated dynamics, *Nature*, **261**, 459–467.

McCulloch, W. S. and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.*, **5**, 115–133.

Misiurewicz, M. (1983), Maps of an interval, in G. Iooss *et al.* (Eds), *Comportement chaothique des systèmes déterministes* pp 567–590 (North-Holland, Amsterdam).

Moon, J. W. (1970), *Counting Labelled Trees*, Canadian Mathematical Monographs No 1 (Alberta).

Muroga, S. (1971), *Threshold Logic and its Application* (Wiley–Interscience, New York).

Nijenhuis, A. and Wilf, H. S. (1975), *Combinatorial Algorithms* (Academic Press, New York).

Nogradi, E. and Labos, E. (1981), Simulations of spontaneous neuronal activity by pseudo-random functions, *Abstracts of Annual Meeting of the Hungarian Physiological Society, Budapest, I.P74*, p 151. (In Hungarian.)

Dynamics of First-Order Partial Differential Equations used to Model Self-Reproducing Cell Populations

P. Brunovský and Jozsef Komorník
Komensky University, Bratislava, CSSR

In this paper we deal with the differential equation

$$u_t + c(x)u_x = f(x,u) \qquad 0 \le x \le 1, \qquad t \ge 0 \tag{1}$$

with the initial condition

$$u(0,x) = v(x) \tag{2}$$

under the following assumptions:

(A1) c is continuously differentiable, $c(0) = 0$, $c(x) > 0$ for $x > 0$.
(A2) f is continuously differentiable and $|f(x,u)| \le k_1 + k_2|u|$ for some $k_1, k_2 > 0$.
(A3) $f(x,0) = 0$ and there exists a $u_0 > 0$ such that $f(0,u_0) \times (u - u_0) < 0$ for $u > 0$, $u \ne u_0$, and $f_u(0,u_0) < 0$.

As we shall see, despite its simplicity this equation exhibits a surprisingly rich dynamics behavior.

The equation has been developed by Lasota in cooperation with Mackey and Ważewska-Czyżewska as a model of the dynamics of a self-reproducing cell population, such as the population of developing red blood cells (erythrocyte precursors). In this model a cell is characterized by a single, scalar variable x to represent maturity, which is normalized to have values in [0, 1]. The state of the population at time t is characterized by a density function $u(t,.)$, i.e.,

$$\int_{x_1}^{x_2} u(t,x)\mathrm{d}x$$

measures the quantity of cells that have a maturity between x_1 and x_2 at time t. Cells proliferate by mitosis, but it is understood that a cell divides into

cells of the same maturity, although any cell can increase its maturity at any stage of the mitotic cycle (*cf.* Mackey and Dörmer, 1982).

The dynamics of the population is characterized by the maturation rate $c(x)$ and the proliferation rate $r(u)$. We assume that c satisfies (A1) above, r is C^{-2}, decreasing, and satisfies $r(0) > c'(0)$, $r'(u) < \rho < 0$ for $u \geq 0$. The balance equation under these assumptions is

$$u_t + (cu)_x = r(u)u$$

By differentiation we obtain equation (1) with $f(x,u) = [r(u) - c'(x)]u$, so f satisfies (A3) with $u_0 = r^{-1}[c(0)]$.

Equations (1) and (2) can be solved using the same characteristics as in the solutions $x = \varphi(t,\xi)$ for

$$dx/dt = c(x) \qquad \varphi(0,\xi) = \xi$$

that is, $u(t,x)$ is a solution of equations (1) and (2) if and only if it satisfies equation (2) and

$$(du/dt)[t,\varphi(t,\xi)] = f\{\varphi(t,\xi), u[t,\varphi(t,\xi)]\} \tag{3}$$

In other words, the graph of u is the forward integral manifold of the system

$$dx/dt = c(x), \qquad du/dt = f(x,u) \tag{4}$$

through the graph of v. From these observations it follows that equations (1) and (2) define a semiflow S on $C^+[0,1]$ – the set of nonnegative continuous functions on [0,1] defined by $S_t(v)(x) = u(t,x)$ – *cf.* Lasota (1981). Here we understand the solution of equations (1) and (2), in a generalized sense, as a uniform limit of classical solutions.

To understand the dynamics of S it is useful to notice that

$$\varphi(t,0) = 0 \tag{5}$$

while $\varphi(t,\xi)$ for any $\xi > 0$ is strictly increasing and eventually leaves [0,1]. Moreover, for fixed $x > 0$ the time that the characteristics through (t,x) spend in a given neighborhood of 0 tends to ∞ with $t \to \infty$, while the time they need to pass from this neighborhood to x remains bounded. Consequently, one can expect that the dynamics of S will largely be determined by the dynamics of the zero characteristics, which are governed by the scalar ordinary differential equation

$$dy/dt = f(0,y) \tag{6}$$

This equation has two stationary points, namely 0 (repellor) and u_0 (attractor), the domain of attraction of the latter being $(0,\infty)$. As an immediate consequence we obtain that a stationary point w of S has to satisfy $w(0) + u_0$ or $w(0) = 0$. The first case is settled by *Theorem 1*.

Theorem 1 [Lasota (1981), Brunovský and Komorník (to appear)]

There exists a unique stationary solution w_0 of equation (1) such that $w_0(0) = u_0$. This solution is asymptotically stable and for each $v \in C^+[0,1]$, such that $v(0) = 0$, and for each $\alpha < f_u(0,u_0)$ there exists a $K > 0$ such that

$$|S_t(v) - w_0| \leq Ke^{-\alpha t}$$

A short proof of *Theorem 1* can be obtained by observing that w_0 is the graph of the (unique) center-unstable manifold $W^u(0,u_0)$ of the point $(0,u_0)$ for system (4).

Turning our attention to the stationary point 0 of equation (6) we note that the set

$$V = \{v \in C^+[0,1] \mid v(0) = 0\}$$

is invariant under S. Also, for each $v \in V$, $S_t(v)$ enters W in finite time where

$$W = \{v \in V \mid 0 \leq v(x) \leq w_0(x) \text{ for all } x\}$$

This is because for a sufficiently large t all the characteristics of $\varphi(t,\xi)$, such that $v(\xi) \geq w_0(\xi)$, have left $[0, 1]$.

Thus, we are led to study the semiflow S on W. *Theorem 2* shows that it is as chaotic as one can imagine.

Theorem 2 [Lasota (1981), Brunovský (1983), Brunovský and Kormorník (1984)]

(1) $S|_W$ is topologically transitive, i.e., admits a dense trajectory.
(2) Each trajectory of $S|_W$ is unstable.
(3) $S|_W$ admits a continuum of periodic points of any period and the set of periodic points is dense in W.
(4) $S|_W$ admits a regular, nontrivial ergodic measure.
(5) $S|_W$ is exact, i.e., there exists a regular, nontrivial probabilistic measure μ on W such that

$$\lim_{t \to \infty} \mu[S_t(A)] = 0$$

as soon as $\mu(A) > 0$.

(6) The topological entropy of $S|_W$ is infinite.

In (4) and (5) by nontrivial we understand that the measure of the set of periodic points has zero measure.

The proof employs the conjugacy F of S to the left-shift T on $C^+[0,\infty)$ defined by

$$F(v)(t) = S_t(v)(1)$$

The following properties of F are easily established:

- $F\colon C^+[0,1] \to C^+[0,\infty)$ is one-to-one and continuous.
- $F(W)$ contains each $g \in C^+[0,\infty)$ which satisfies

$$g(t) - w^0(1) \le Ke^{-\alpha t}$$

for some $K > 0$ and some $\alpha < f_u(0,u_0)$.

- $F \circ S_t = T_t \circ F$ where T is defined by

$$T_t(g)(s) = g(s + t) \qquad t \ge 0, \qquad s \ge 0$$

The conjugacy F makes it possible to study T instead of S, which is much simpler, and then carry the results back to S. For instance, any nonnegative periodic function g is a periodic point of T. If $0 \le g(t) < w_0(1)$ for all t then $g \in F(W)$ so $F^{-1}(g) \in W$ is a periodic point of S of the same period as g. The proof of the rest of (1) to (3) in *Theorem 2* proceeds along the same lines, the details being somewhat more complicated.

The proof of (4) and (5) of *Theorem 2* requires a more extensive machinery. An easy argument shows that (5) implies (4). To establish (5) we first construct a nontrivial exact measure on $C[0,\infty)$ endowed by the topology of almost uniform convergence.

The space $C[0,\infty)$ is metricizable and admits a metric, each open ball of which is a countable intersection of measurable cylinders. This is why its Borel σ-algebra coincides with its Kolmogorov σ-algebra generated by measurable cylinders. We define the exact measure m as the unique measure generated by the Gaussian stationary process Y with the realizations $Y_t(g) = g(t)$, $g \in C[0,\infty)$, and the autocovariance function

$$\operatorname{cov}(Y_t, Y_s) = \max\{1 - |t - s|, 0\}$$

Since $C[0,\infty)$ is a Polish space, m can be extended to a σ-algebra in which continuous images and preimages of measurable sets are measurable, thus $T_t(A)$ is measurable if A is. To see why m is exact, note that events for which

the time distance exceeds 1 are independent. It is clear that $m[T_t(C)] = 1$ for any measurable cylinder C that is constrained at time instants not exceeding t.

The independence of distant events also implies that the set of bounded (and, in particular, periodic) functions has zero measure, so m is nontrivial. This, however, has the disadvantage that we cannot define the exact measuregp on W by the formula

$$\mu(A) = \mu[F(A)] / \mu[F(W)]$$

since $F(W)$ is bounded and $F(W) = 0$.

In order to carry m back to W by F we have to perform a preliminary scaling of the real line. This is possible due to the following *Lemma*.

Lemma [Brunovský and Komorník (to appear)]

There exists a $K > 0$ such that $\mu(M) > 0$ where

$$M = \{g \in C[0,\infty) \mid \Phi^{-1}(K_t^{-7}) \leq g(t) \leq PHIgp^{-1}(1 - Kt^{-7})\}$$

where Φ is the distribution function of the normalized normal distribution.

Now, for $C_0 = \{g \in C[0,\infty) \mid 0 < g(t) < w_0(1)\}$ we define $H\colon\ C_0 \to C[0,\infty)$ by

$$H(g)(t) = h[g(t)]$$

where

$$h(y) = \bar{\Phi}^{-1}[y / w^0(1)]$$

For $A \subset C_0$ we define

$$m_0(A) = m[H(A)]$$

Then, we have $m_0[H^{-1}(M)] > 0$. Since $F(W) \cap C_0 \subseteq H^{-1}(M)$, we have

$$m_0[F(W) \cap C_0] \geq m_0[\{g \mid Lt^{-7} \leq g(t) \leq w_0(1) \mid -Lt^{-7}\}] > 0$$

where $L = Kw_0(1)$, so we can define for $A \subset W$

$$\mu(A) = m_0[F(A) \cap D] / m_0[F(W) \cap D]$$

Then, μ is a measure called for by (4) and (5) of *Theorem 2*.

To establish (6) one notes that the shift on C_0 contains the Bernoulli shift, with any finite number of symbols as its restriction. This observation is due to K. Sigmund.

We remark that physiologically w_0 can be interpreted as the stable equilibrium density of the population under normal conditions, while the chaotic behavior of S on W can be interpreted as proliferative disorders caused by a lack of the least-mature (stem) cells.

Forgetting about the physiological context for a while we can consider equations (1) and (2) with f, instead of satisfying (A3), being defined on the entire real line, with $f(0,u)$ having only simple zeros and being bounded away from zero at infinity for cases when its zeros are bounded. The proofs of *Theorems 1* and *2* can be adapted in a straightforward way to prove *Theorem 3*.

Theorem 3

(1) If u_0 is a stable equilibrium of equation (6), then there is a unique stationary solution of equation (1) such that $w_0(0) = u_0$.
(2) The domain of attraction of w_0 consists of those $v \in C[0,1]$ that have $v(0)$ in the domain of attraction of u_0.
(3) If u is an unstable equilibrium of equation (6), then the set

$$W = \{v \mid v(0) = \bar{u}, [x, v(x)] \in W^u(0,\bar{u}) \text{ for } x > 0\}$$

where $W^u(0,\bar{u})$ is the unstable manifold of $(0,\bar{u})$ for equation (4)] is attracting in $V = \{v \mid v(0) = \bar{u}\}$ and $S|_W$ is chaotic in the sense of *Theorem 2*.

We note that the set W is attracting in V, but not in the entire space.

Considering interacting populations, such as populations of developing blood cells of different kinds, we are led to study systems of equations such as

$$u_{it} + c_i(x)u_{ix} = f(x,u) \qquad i = 1, \ldots, n, \qquad t > 0, \qquad 0 \le x \le 1$$
$$u_i(0,x) = v_i(x)$$

or, in vector form,

$$u_t + C(x)u_x = f(x,u) \tag{7}$$
$$u(0,x) = v(x) \tag{8}$$

where $C(x) = \text{diag}\{c_1(x), \ldots, c_n(x)\}$. We assume that all c_i satisfy (A1) and f satisfies (A2).

Integrating the equations along characteristics (which may, in general, be different for different components) and using successive approximation techniques it is not difficult to prove that equation (7) defines a semiflow S on the set of continuous functions on [0,1] with values in R^n (to be also denoted by $C[0,1]$).

First, consider the case of equal c_is. Under the assumption that equation (6) is a generic gradient system (by generic we mean that its stationary points are hyperbolic and their stable and unstable manifolds intersect transversally) *Theorem 3* remains valid for any n.

However, physiologically more interesting is the case of unequal c_is. Lasota conjectured that proliferation disorders can be caused by the occurrence of large differences in maturation rates.

This case is mathematically more difficult and only fragments of a theory are available. Nevertheless, it is clear that *Theorem 3* does not extend, in particular (1) does not hold in the higher dimensions.

Let u_0 be a stable equilibrium of equation (6), i.e., $A = f_u(0,u_0)$ has all its eigenvalues in the open, left halfplane and let

$$K = \mathrm{diag}\{c'_1(0), \ldots, c'_n(0)\}, \qquad W_0 = \{v \in C[0,1] \mid v(0) = u_0\}$$

Then, W_0 is attracting, but may not consist of a single solution. In order that W_0 contains a single stationary solution it is necessary that the spectrum of $K^{-1}A$ is contained in the closed, left halfplane and sufficient that it is contained in the open left one. It can be shown that this unique stationary solution w_0 is asymptotically stable if the linear semigroup generated by the equation

$$z_t = C(x)z_x + f_u[x, w_0(x)]z \tag{9}$$

on the space of those $v \in C[0,1]$ that satisfy $z(0) = 0$ isgp asymptotically stable. The spectrum of the generator of this semigroup is the set of those λ for which

$$L(\lambda) = K^{-1}A - \lambda K^{-1}$$

has an eigenvalue in the closed, right halfplane. Therefore, in order that w_0 is asymptotically stable it is necessary that for all $\lambda \geq 0$ the spectrum of $L(\lambda)$ lies in the open left halfplane – *cf.* Pazy (1983). Unfortunately, since equation (9) has the character of a hyperbolic partial differential equation, linear semigroup theory does not provide for a sufficient condition for its asymptotic stability in terms of the spectrum of its infinitesimal generator. Therefore, the problem of finding efficient criteria for asymptotic stability of w_0 remains open.

Supporting Lasota's conjecture, we note that in this case A is a stable matrix, but $K^{-1}A$ is not. For the linear equation

$$u_t + Ku_x = Au$$

the set W_0 for $u_0 = 0$ is attracting, but each solution of W_0 is unstable.

References

Brunovský, P. (1983), Notes on chaos in the cell population partial differential equation, *Nonlinear Analysis*, **7**, 167–176.

Brunovský, P. and Komorník, J. (1984), Ergodicity and exactness of the shift on $C[0,\infty)$ and the semiflow of a first order partial differential equation, *Journal of Mathematical Analysis and Applications*, **104**, 235–245.

Brunovský, P. and Komorník, J. (to appear), Explicit definition of an exact measure for the semiflow of a first order partial differential equation.

Lasota, A. (1981), Stable and chaotic solutions of a first order partial differential equation, *Nonlinear Analysis*, **5**, 1181–1193.

Mackey, M.C. and Dörmer, P. (1982), Continuous maturation of proliferating erythroid precursors, *Cell Tissue Kinetics*, **15**, 381–392.

Pazy, A. (1983), *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer, New York).

THE INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS

is a nongovernmental research institution, bringing together scientists from around the world to work on problems of common concern. Situated in Laxenburg, Austria, IIASA was founded in October 1972 by the academies of science and equivalent organizations of twelve countries. Its founders gave IIASA a unique position outside national, disciplinary, and institutional boundaries so that it might take the broadest possible view in pursuing its objectives:

To promote international cooperation in solving problems arising from social, economic, technological, and environmental change

To create a network of institutions in the national member organization countries and elsewhere for joint scientific research

To develop and formalize systems analysis and the sciences contributing to it, and promote the use of analytical techniques needed to evaluate and address complex problems

To inform policy advisors and decision makers about the potential application of the Institute's work to such problems

The Institute now has national member organizations in the following countries:

Austria
The Austrian Academy of Sciences

Bulgaria
The National Committee for Applied Systems Analysis and Management

Canada
The Canadian Committee for IIASA

Czechoslovakia
The Committee for IIASA of the Czechoslovak Socialist Republic

Finland
The Finnish Committee for IIASA

France
The French Association for the Development of Systems Analysis

German Democratic Republic
The Academy of Sciences of the German Democratic Republic

Federal Republic of Germany
Association for the Advancement of IIASA

Hungary
The Hungarian Committee for Applied Systems Analysis

Italy
The National Research Council

Japan
The Japan Committee for IIASA

Netherlands
The Foundation IIASA–Netherlands

Poland
The Polish Academy of Sciences

Sweden
The Swedish Council for Planning and Coordination of Research

Union of Soviet Socialist Republics
The Academy of Sciences of the Union of Soviet Socialist Republics

United States of America
The American Academy of Arts and Sciences

Vol. 238: W. Domschke, A. Drexl, Location and Layout Planning. IV, 134 pages. 1985.

Vol. 239: Microeconomic Models of Housing Markets. Edited by K. Stahl. VII, 197 pages. 1985.

Vol. 240: Contributions to Operations Research. Proceedings, 1984. Edited by K. Neumann and D. Pallaschke. V, 190 pages. 1985.

Vol. 241: U. Wittmann, Das Konzept rationaler Preiserwartungen. XI, 310 Seiten. 1985.

Vol. 242: Decision Making with Multiple Objectives. Proceedings, 1984. Edited by Y.Y. Haimes and V. Chankong. XI, 571 pages. 1985.

Vol. 243: Integer Programming and Related Areas. A Classified Bibliography 1981–1984. Edited by R. von Randow. XX, 386 pages. 1985.

Vol. 244: Advances in Equilibrium Theory. Proceedings, 1984. Edited by C.D. Aliprantis, O. Burkinshaw and N.J. Rothman. II, 235 pages. 1985.

Vol. 245: J.E.M. Wilhelm, Arbitrage Theory. VII, 114 pages. 1985.

Vol. 246: P.W. Otter, Dynamic Feature Space Modelling, Filtering and Self-Tuning Control of Stochastic Systems. XIV, 177 pages. 1985.

Vol. 247: Optimization and Discrete Choice in Urban Systems. Proceedings, 1983. Edited by B.G. Hutchinson, P. Nijkamp and M. Batty. VI, 371 pages. 1985.

Vol. 248: Plural Rationality and Interactive Decision Processes. Proceedings, 1984. Edited by M. Grauer, M. Thompson and A.P. Wierzbicki. VI, 354 pages. 1985.

Vol. 249: Spatial Price Equilibrium: Advances in Theory, Computation and Application. Proceedings, 1984. Edited by P. T. Harker. VII, 277 pages. 1985.

Vol. 250: M. Roubens, Ph. Vincke, Preference Modelling. VIII, 94 pages. 1985.

Vol. 251: Input-Output Modeling. Proceedings, 1984. Edited by A. Smyshlyaev. VI, 261 pages. 1985.

Vol. 252: A. Birolini, On the Use of Stochastic Processes in Modeling Reliability Problems. VI, 105 pages. 1985.

Vol. 253: C. Withagen, Economic Theory and International Trade in Natural Exhaustible Resources. VI, 172 pages. 1985.

Vol. 254: S. Müller, Arbitrage Pricing of Contingent Claims. VIII, 151 pages. 1985.

Vol. 255: Nondifferentiable Optimization: Motivations and Applications. Proceedings, 1984. Edited by V.F. Demyanov and D. Pallaschke. VI, 350 pages. 1985.

Vol. 256: Convexity and Duality in Optimization. Proceedings, 1984. Edited by J. Ponstein. V, 142 pages. 1985.

Vol. 257: Dynamics of Macrosystems. Proceedings, 1984. Edited by J.-P. Aubin, D. Saari and K. Sigmund. VI, 280 pages. 1985.

Vol. 258: H. Funke, Eine allgemeine Theorie der Polypol- und Oligopolpreisbildung. III, 237 pages. 1985.

Vol. 259: Infinite Programming. Proceedings, 1984. Edited by E.J. Anderson and A.B. Philpott. XIV, 244 pages. 1985.

Vol. 260: H.-J. Kruse, Degeneracy Graphs and the Neighbourhood Problem. VIII, 128 pages. 1986.

Vol. 261: Th.R. Gulledge, Jr., N.K. Womer, The Economics of Made-to-Order Production. VI, 134 pages. 1986.

Vol. 262: H.U. Buhl, A Neo-Classical Theory of Distribution and Wealth. V, 146 pages. 1986.

Vol. 263: M. Schäfer, Resource Extraction and Market Structure. XI, 154 pages. 1986.

Vol. 264: Models of Economic Dynamics. Proceedings, 1983. Edited by H.F. Sonnenschein. VII, 212 pages. 1986.

Vol. 265: Dynamic Games and Applications in Economics. Edited by T. Başar. IX, 288 pages. 1986.

Vol. 266: Multi-Stage Production Planning and Inventory Control. Edited by S. Axsäter, Ch. Schneeweiss and E. Silver. V, 264 pages. 1986.

Vol. 267: R. Bemelmans, The Capacity Aspect of Inventories. IX, 165 pages. 1986.

Vol. 268: V. Firchau, Information Evaluation in Capital Markets. VII, 103 pages. 1986.

Vol. 269: A. Borglin, H. Keiding, Optimality in Infinite Horizon Economies. VI, 180 pages. 1986.

Vol. 270: Technological Change, Employment and Spatial Dynamics. Proceedings 1985. Edited by P. Nijkamp. VII, 466 pages. 1986.

Vol. 271: C. Hildreth, The Cowles Commission in Chicago, 1939–1955. V, 176 pages. 1986.

Vol. 272: G. Clemenz, Credit Markets with Asymmetric Information. VIII, 212 pages. 1986.

Vol. 273: Large-Scale Modelling and Interactive Decision Analysis. Proceedings, 1985. Edited by G. Fandel, M. Grauer, A. Kurzhanski and A.P. Wierzbicki. VII, 363 pages. 1986.

Vol. 274: W.K. Klein Haneveld, Duality in Stochastic Linear and Dynamic Programming. VII, 295 pages. 1986.

Vol. 275: Competition, Instability, and Nonlinear Cycles. Proceedings, 1985. Edited by W. Semmler. XII, 340 pages. 1986.

Vol. 276: M.R. Baye, D.A. Black, Consumer Behavior, Cost of Living Measures, and the Income Tax. VII, 119 pages. 1986.

Vol. 277: Studies in Austrian Capital Theory, Investment and Time. Edited by M. Faber. VI, 317 pages. 1986.

Vol. 278: W.E. Diewert, The Measurement of the Economic Benefits of Infrastructure Services. V, 202 pages. 1986.

Vol. 279: H.-J. Büttler, G. Frei and B. Schips, Estimation of Disequilibrium Models. VI, 114 pages. 1986.

Vol. 280: H.T. Lau, Combinatorial Heuristic Algorithms with FORTRAN. VII, 126 pages. 1986.

Vol. 281: Ch.-L. Hwang, M.-J. Lin, Group Decision Making under Multiple Criteria. XI, 400 pages. 1987.

Vol. 282: K. Schittkowski, More Test Examples for Nonlinear Programming Codes. V, 261 pages. 1987.

Vol. 283: G. Gabisch, H.-W. Lorenz, Business Cycle Theory. VII, 229 pages. 1987.

Vol. 284: H. Lütkepohl, Forecasting Aggregated Vector ARMA Processes. X, 323 pages. 1987.

Vol. 285: Toward Interactive and Intelligent Decision Support Systems. Volume 1. Proceedings, 1986. Edited by Y. Sawaragi, K. Inoue and H. Nakayama. XII, 445 pages. 1987.

Vol. 286: Toward Interactive and Intelligent Decision Support Systems. Volume 2. Proceedings, 1986. Edited by Y. Sawaragi, K. Inoue and H. Nakayama. XII, 450 pages. 1987.

Vol. 287: Dynamical Systems. Proceedings, 1985. Edited by A.B. Kurzhanski and K. Sigmund. VI, 215 pages. 1987.

This series reports new developments in mathematical economics, economic theory, econometrics, operations research, and mathematical systems, research and teaching – quickly, informally and at a high level. The type of material considered for publication includes:

1. Preliminary drafts of original papers and monographs
2. Lectures on a new field or presentations of a new angle in a classical field
3. Seminar work-outs
4. Reports of meetings, provided they are
   a) of exceptional interest and
   b) devoted to a single topic.

Texts which are out of print but still in demand may also be considered if they fall within these categories.

The timeliness of a manuscript is more important than its form, which may be unfinished or tentative. Thus, in some instances, proofs may be merely outlined and results presented which have been or will later be published elsewhere. If possible, a subject index should be included. Publication of Lecture Notes is intended as a service to the international scientific community, in that a commercial publisher, Springer-Verlag, can offer a wide distribution of documents which would otherwise have a restricted readership. Once published and copyrighted, they can be documented in the scientific literature.

Manuscripts

Manuscripts should be no less than 100 and preferably no more than 500 pages in length. On request, the publisher will supply special paper with the typing area outlined and essentials for the preparation of camera-ready manuscripts. Manuscripts should be sent directly to Springer-Verlag Heidelberg or Springer-Verlag New York.

Springer-Verlag, Heidelberger Platz 3, D-1000 Berlin 33
Springer-Verlag, Tiergartenstraße 17, D-6900 Heidelberg 1
Springer-Verlag, 175 Fifth Avenue, New York, NY 10010/USA
Springer-Verlag, 37-3, Hongo 3-chome, Bunkyo-ku, Tokyo 113, Japan

ISBN 3-540-17698-5
ISBN 0-387-17698-5