Foundations of Catastrophe Theory

Antal Majthay

Foundations of Catastrophe Theory

Antal Majthay
Graduate School of Business Administration
University of Florida

Pitman Advanced Publishing Program
BOSTON · LONDON · MELBOURNE

PITMAN PUBLISHING INC
1020 Plain Street, Marshfield, Massachusetts 02050

PITMAN PUBLISHING LIMITED
128 Long Acre, London WC2E 9AN

Associated Companies
Pitman Publishing Pty Ltd, Melbourne
Pitman Publishing New Zealand Ltd, Wellington
Copp Clark Pitman, Toronto

First published 1985

AMS Subject Classifications: (main) 5802, 58C25, 58C27, 58C28
(subsidiary) 57RXX

ISSN 8756-7415

Library of Congress Cataloging in Publication Data

Majthay, Antal.
Foundations of catastrophe theory.

Bibliography: p.
Includes index.
1. Catastrophes (Mathematics) I. Title.
QA614.58.M35 1985 514'.74 85-6256
ISBN 0-273-08551-4

British Library Cataloguing in Publication Data

Majthay, Antal
Foundations of catastrophe theory.—(Advanced texts and surveys in pure and applied mathematics, ISSN 8750 7415: 26)
1. Catastrophes (Mathematics)
I. Title II. Series
514'.72 QA614.58

ISBN 0-273-08551-4

Filmset and printed in Northern Ireland by
The Universities Press (Belfast) Ltd

Preface

During the past few decades the increasing complexity of the economy and the rapid growth of technology have put ever-increasing demands on mathematics. New problems have required solution. Many such problems were amenable to mathematical formulation, but fewer and fewer could be solved in closed form. The traditional notions about solutions of mathematical models had to be re-evaluated.

The ready availability of computers offered an easy way out of this difficulty. Problems were solved numerically. An explosive development of numerical methodology ensued. After a few decades of this development it became clear, however, that numerical solutions cannot always be considered satisfactory, no matter how accurate and fast. A numerical solution is only a set of numbers applicable to one particular set of numerically specified conditions. A set of numbers may fail to give insight into a situation: for instance, numerically obtained solutions cannot tell whether or not the response of a system under study is, say, continuous or stable. When a model cannot be solved with all possible boundary conditions, the possibility of some surprise cannot be ruled out.

It is necessary to complement numerical mathematics by a methodology capable of providing the missing insight into the nature, stability and sensitivity of numerically obtained solutions. The ideas utilized for this purpose were borrowed and combined from several hitherto abstruse branches of mathematics.

Some of the results obtained from this new line of research have been amazing. It was discovered, for instance, that, as a rule, the stability of a system breaks down only according to certain relatively simple patterns that can be described explicitly. This is the central result of elementary catastrophe theory. It was also discovered that a well-determined system may respond to very regular stimuli in a totally chaotic manner. The comfortable conceptual dichotomy, deterministic versus stochastic, was severely shaken.

This book grew out of two seminars on stability, singularities and catastrophes. The participants were faculty and doctoral students in engineering and the social sciences. Before any substantive issues could be discussed, a large amount of background material had to be covered: the book discusses subjects in which the existing literature was inadequate, for one reason or another.

The book is aimed at the well educated nonspecialist who has a strong motivation. Since the prospective reader may not have the benefit of a general background context for the theory, I differ from the usual style of writing mathematics in laying heavy emphasis on motivation. Although certain specialists may disapprove of such an approach, an example illustrates my reasons: I have included a detailed systematic analysis of the canonical elementary catastrophes. An elementary catastrophe is a family of real-valued functions that is specified only in a rather vague, topological sense. It may seem, therefore, pointless to carry out a detailed classical analysis of an arbitrarily chosen particular representative of the family. Yet, after several different attempts, including the use of computer-generated pictures, this was the only method I could find to elicit the "Aha!" reaction from my audience. What may appear to be false starts and chatty explanations are included in several places for similar reasons. Otherwise I have tried to keep a gradual movement from the simple toward the more abstract. Each chapter is closed by a terse and rigorous summary of the main ideas and by a short guide to the literature.

Gainesville, Florida — A.M.
May 1985

Contents

Introduction

Catastrophe theory is a branch of mathematics. It grows where algebra, calculus, and topology meet each other, and is concerned with the study of real-valued functions of several real variables.

If some partial derivative of a real-valued function is non-zero at some point of its domain, the behavior of the function near this point can safely be predicted from this fact. The linear terms of the Taylor expansion determine the qualitative behavior of the function. Such points are said to be *regular.*

If all partial derivatives of the function are zero at some point, almost anything can happen. The function may have a minimum, a maximum, a saddle, or it can be even more complicated. Such points are called *critical* points.

Is it possible to describe all varieties of behavior at a critical point? Since the answer is negative, we must be content with something less. In a sense, catastrophe theory was conceived when this question was first asked in a different way.

Let us suppose that a real-valued function of several variables is perturbed slightly. What can be said about the result? It turns out that a regular point will remain regular. If the function has a non-degenerate minimum, maximum, or saddle, the perturbed function will again have a minimum, a maximum, or a saddle close to the original one. If a function has a critical point that is none of the above, the outcome depends on the nature of that critical point and on the particular perturbation. Such critical points are called *degenerate.*

Suppose now that we wish to study this case. We cannot reasonably ask what could happen when a function with a degenerate critical point is perturbed since, as it was just mentioned, anything could happen. We can however ask what will typically happen.

Catastrophe theory in the strict sense, more properly called *elementary* catastrophe theory, specifies the meaning of 'typical' and then sets out to answer this question. Since even this is too ambitious, a complete answer is not expected. Instead we find complete answers for specialized versions of the question.

Searching for the outcome of a perturbation is not a purely mathematical pastime. Functions are used to model natural or social phenomena. A model is useful only if it reflects some observable phenomenon. An observable phenomenon has to persist for a finite time interval. Since any physical, biological, or social system interacts constantly with its environment, it is always subject to perturbations of some kind. If a phenomenon persists for some time interval, no matter how short this interval is, it has to be stable in some sense. Any mathematical model of the phenomenon, such as a function, should reflect this ability to withstand small perturbations. It should also be stable.

Sudden drastic transitions of a physical, biological, or social system from one state to another are of much interest. The state before the transition as well as the state after the transition persist, but the act of the transition is ephemeral. The observer will notice the fact that the transition has occurred but it is virtually impossible to freeze the system in the state of transition. In other words, the states before and after the transition are stable, but the state in between is not. If it were, it would be a third stable state and we would be forced to investigate two transitions.

Any mathematical model of the state before the transition should be stable. The same is true for any model of the state after the transition. What about the instant of the transition? Since it is ephemeral, unstable, it should be modeled by an unstable mathematical object, such as an unstable function. This unstable function should, however, be somehow related to the two stable ones, to the models of the two stable states before and after the transition. When the model of the transition is perturbed, it should change to either the one or the other of those two stable models.

The next natural question is: how to model the phenomenon of the transition itself. The model should be some mathematical object that contains all the above, and more. The transition from one state to another has been observed. In this sense it is a stable phenomenon, and its mathematical model should be stable.

When we attempt to unite all these requirements into a useful mathematical notion, the result is a family of functions. This family is the model of the transition from one state to another. The members of the family fall into three distinct groups. Those in the first group are stable and they are all very much alike. They represent the state of the system in its various phases of evolution before the transition. Those in the second group are also stable and they are also alike, but differ from the members of the first group. They represent the state of the system in its various phases of evolution after the transition. The two groups are separated by an unstable member that is unstable because it bears all the marks of both groups. This unstable member represents the instance of transition from one state to the other. Although the family contains unstable elements, it should be stable as a whole since it represents an observable phenomenon, the state transition.

From this point of view it is not enough to study functions with degenerate critical points. They should be studied in context. Embed them into a family of functions so that they are surrounded by stable neighbors and in such a manner that the resulting family as a unit is stable. Such families would then be the appropriate models for state transitions.

A sudden change of the state of a system may be catastrophic for the participants in the change. From this observation it is only a short step to call all such changes catastrophes, especially when the term is coined in an environment with a highly developed sense for dramatic expressions. If so, it is only logical to also call the mathematical models representing these changes by the same name. Thus we come to the mathematical notion of a catastrophe: a stable family of functions that contains unstable members.

The study of such families uses tools that are unfamiliar to all but a few experts. To make things even more difficult, even familiar tools are sometimes used in unfamiliar ways. This has created a wide gap between the few who understand, and the multitude who do not.

The literature responded to this situation with books that attempt to convey the results, without really bridging the gap. This book is an attempt at bridging the gap.

The reader is assumed to be college educated with the usual two or three calculus classes and some sprinkling of other mathematical background. In addition, he should be intellectually curious and possess some patience for digesting unfamiliar concepts. The book will build a bridge from calculus classes to the better quality catastrophe theory literature. As an added bonus, the reader will learn about several subjects that are omitted from most textbooks and college courses. These subjects are either not taught yet, or else they are simply assumed to be known already.

As is always the case in scientific inquiry, the search for pertinent notions and methods leads to new problems. The newly developed methods can be used to answer new questions, the new notions start a life of their own. After a while it is hard to tell where the answer to the original question ends and where some new branch of inquiry has just started. So it is hard to tell where the boundaries of catastrophe theory are.

Catastrophe theory means different things to different authors. To the reader of this book it will mean a branch of mathematics. To readers of popular science magazines it means fancifully curved surfaces with catchy names, like cusp, butterfly and umbilic. They seem to somehow explain all sorts of surprising events, like the collapse of the stock market, the outbreak of war, or the biting of dogs.

From the viewpoint of its founder, René Thom, catastrophe theory is not a specified body of knowledge but rather a scientific program. It seeks to find explanations to the variety and evolution of forms in nature. A theory of morphogenesis. '. . . many phenomena of common experience, in themselves trivial (often to the point that they escape attention altogether!)—for example, the cracks in an old wall, the shape of a cloud, the path of a falling leaf, or the froth on a pint of beer—are very difficult to formalize, but is it not possible that a mathematical theory launched for such homely phenomena might, in the end, be more profitable for science?' Thom (1975).

A few more quotations should illustrate the range of views about the nature of catastrophe theory.

'Catastrophe theory is a new mathematical method for describing the evolution of forms in nature.' Zeeman (1976)

'Catastrophe theory is a controversial new way of thinking about change . . .' Woodcock and Davis (1978)

'As a part of mathematics, catastrophe theory is a theory about singularities. When applied to scientific problems, therefore, it deals with the properties of discontinuities directly, without reference to any specific underlying mechanism.' Saunders (1980)

'Catastrophe theory attempts to study how the qualitative nature of the solutions of equations depends on the parameters that appear in the equations.' . . . 'Elementary catastrophe theory is the study of how the equilibria $\psi_j(C_\alpha)$ of $V(\psi_j; C_\alpha)$ change as the control parameters C_α change.' Gilmore (1981)

1 Coordinate systems

One characteristic feature of catastrophe theory is the free use of all kinds of non-Cartesian coordinate systems. Very often, the specific form of the system is not described. Statements like this are common: 'There is a coordinate system such that . . .'. For this reason our first task is to extend the notion of a coordinate system, so that we can feel comfortable with a more relaxed use of it.

A coordinate system is a labeling system. Each element in a set of objects is assigned a label and distinct elements receive distinct labels. The labels are usually ordered n-tuples of real or complex numbers. The sets to be labeled by coordinates are lines, surfaces, spaces.

When the term coordinate system is mentioned most of us will, almost automatically, think of two or three mutually perpendicular straight lines with some points marked on them. We think of a Cartesian system. The simplicity of the Cartesian system is so successful that it pervades most of our ideas.

The widespread use of the Cartesian system is motivated by its mathematical convenience rather than by some inherent property of the problems for which it is used. A resource allocation problem, for instance, may deal with various activities that are related to each other only through their use of common resources. Yet it is customary to treat these activities as if an intrinsic geometric relationship existed between them, as if they were mutually perpendicular to each other; that is, the problem is exhibited in a Cartesian coordinate system. This is done for convenience, and for lack of a better alternative.

Compare the resource allocation problem with that of satellite motion or the motion of a pendulum. When studying circular type motions the analyst is almost forced to use a polar, or a cylindrical coordinate system.

The inadequacy of a coordinate system is often revealed by a need to change it: well-known examples are the sequence of coordinate changes required to find the optimal solution to a linear program or to diagonalize a quadratic form. The technique of integration by substitution is nothing but a skillful change of coordinates which is forced on us by the nature of the integrand.

1.1 Coordinates on a straight line

The usual method of assigning coordinates to a straight line is well known. Two distinct points are chosen on the line, the number 0 is arbitrarily assigned to one of them and the number 1 to the other. Then any arbitrary point p is assigned a real number by comparing the length and direction of the segment $[0, p]$ with the length and direction of segment $[0, 1]$. This is a simple and useful but rigid rule.

Far more flexible rules can be obtained by realizing that the major purpose of coordinates is identification: namely, assigning a name to each and every point such that distinct points have distinct names. With this general notion in mind we come to the following, tentative definition:

A coordinate system on a straight line L is an injective map $\phi: L \to R$ from the line L into the set of real numbers R. Every point p on the line L has a coordinate $\phi(p) \in R$, and distinct points have distinct coordinates. The map ϕ is variously called a *coordinate map*, a *coordinate function*, a *coordinate system*, or a *chart map*. The inverse map $\phi^{-1}: \phi(L) \to L$ is called a *parametrization* of the line L.

It is helpful initially to identify the set R of real numbers with a straight line given coordinates in the above way. With this picture in mind the coordinate function ϕ can be visualized as an injection from the line L to be assigned coordinates into the model line R (Fig. 1.1).

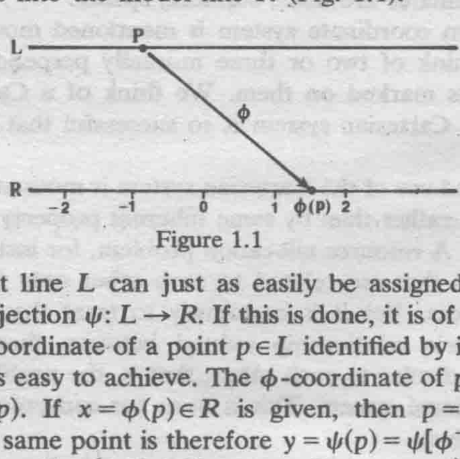

Figure 1.1

The same straight line L can just as easily be assigned coordinates by means of another injection $\psi: L \to R$. If this is done, it is of interest to know how to find the ψ-coordinate of a point $p \in L$ identified by its ϕ-coordinate, or vice-versa. This is easy to achieve. The ϕ-coordinate of p is $\phi(p)$ and the ψ-coordinate is $\psi(p)$. If $x = \phi(p) \in R$ is given, then $p = \phi^{-1}(x)$ and the ψ-coordinate of the same point is therefore $y = \psi(p) = \psi[\phi^{-1}(x)]$. If $y \in R$ is known, we obtain $x = \phi[\psi^{-1}(x)]$. This can easily be seen in Fig. 1.2.

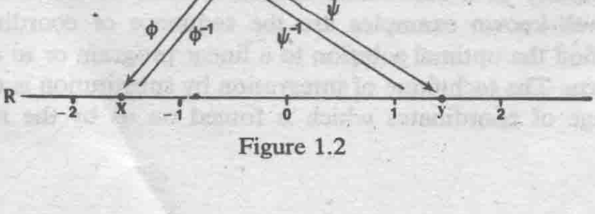

Figure 1.2

A change in coordinates is thus accomplished by the composite maps $h=\psi\circ\phi^{-1}$ and $h^{-1}=\phi\circ\psi^{-1}$, called *transition maps*. $\psi\circ\phi^{-1}$ is read 'the inverse of ϕ followed by ψ', and it is defined by $\psi\circ\phi^{-1}(x)=\psi[\phi^{-1}(x)]$ (see Appendix).

The conceptual picture presented above is quite clear. It helps to clarify several practical procedures. It is not, however, an accurate picture of practical attitudes. In everyday practice it is the coordinates rather than the coordinatized points that play the leading role. Very often L coincides with R and we have to force the above framework by imagining two copies of R with the identity map as the coordinate map between them. Except for mechanics and some other disciplines which study phenomena in the space of our everyday experience, geometric objects like lines are only used as mathematical abstractions. The primary, directly observable, objects in, say, an economic problem are numbers.

It may happen that the way the real numbers naturally enter into some expression is inconvenient. If so, a substitution may solve the difficulty. Integration by substitution is perhaps the best known and most widely used example.

A *substitution* is the replacement of the values of a variable by the values of a bijective (one-to-one and onto) function. A long chain of substitutions may become quite confusing. The interpretation of subsuitutions as coordinate changes is an excellent conceptual framework for avoiding the confusion.

In spite of the fact that the coordinates $x\in R$ and the coordinate changes $h=\psi\circ\phi^{-1}$ are the mathematical concepts of immediate practical interest, the clear conceptual distinction between the set L to be coordinatized and usually thought of as a geometric object, and its model R, together with the notion of the coordinate maps ϕ, ψ, simplify the understanding of many intricate problems.

A simple example may be helpful here. Suppose that for some reason we have to make the following substitution on the real line: $y=x^3+x$. We observe first that the transition map $h\colon R\to R$ defined by $h(x)=x^3+x$ is a strictly increasing, infinitely differentiable function on the whole real line whose derivative $h'(x)=3x^2+1$ is everywhere positive. Since h is strictly increasing, it is bijective. It has an inverse $h^{-1}\colon R\to R$. Its inverse is also infinitely differentiable by the inverse function theorem (see Chapter 3).

One possibility for putting this substitution into the above theoretical framework could be the following. Take three copies of the real line (see Fig. 1.3).

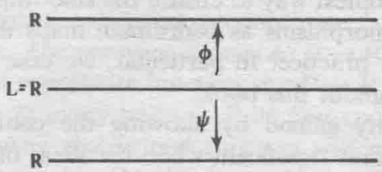

Figure 1.3

The line in the middle is the set L to be coordinatized, the ones on the top and on the bottom are the two models of L. Take the identity map $x=\phi(p)=p$ as the coordinate system $\phi: L \to R$.

For the second coordinate system choose the bijection $\psi: L \to R$ defined by $y=\psi(p)(=h(p))=p^3+p$. The resulting transition map $\psi \circ \phi^{-1}$ is then indeed $h(x)=\psi[\phi^{-1}(x)]=x^3+x$.

In general, any substitution h can be thought of as the composition $\psi \circ \phi^{-1}$ with $\phi=$identity and $\psi=h$. It could just as well be thought of as the composition of $\phi^{-1}=h$ and $\psi=$identity, of course.

The notions of coordinate maps and transition maps, as just defined, are too general to be of any real use. Some additional requirements are usually imposed on these maps. The least to be expected from a coordinate map is that the coordinates of points that are close to each other should also be close to each other and, conversely, if the coordinates are close, the points should also be close. This can be expressed by saying that both the coordinate map and its inverse, the parametrization, should be continuous. A continuous bijection whose inverse is also continuous is called a *homeomorphism*. Coordinate maps should be homeomorphisms.

Since compositions of homeomorphisms are homeomorphisms, transition maps are automatically homeomorphisms. For most purposes this is not enough, however. The real line can be both the domain and the range of functions. Functions may possess desirable properties like differentiability or analyticity. A change of coordinates should not destroy those properties. For this reason any of a whole scale of successively stronger requirements may be imposed on the transition maps: continuous differentiability, twice continuous differentiability, ..., k-times continuous differentiability, ..., infinite, sometimes called indefinite, continuous differentiability, or even analyticity. We may even want the transition map to be an affine linear function or a homogeneous linear function. As may be apparent from the context, in each of these cases the inverse transition maps should possess the same property.

A differentiable map with a differentiable inverse is called a *diffeomorphism*. This term is used in a variety of senses depending on the degree of differentiability required. As was mentioned in the previous paragraph, the requirement can vary from once continuous differentiability to analyticity. An overworked, but expressive, colloquialism very often used here is 'smooth'. It may mean anything from once continuously differentiable to indefinitely continuously differentiable, depending on the context.

In case $L=R$, the simplest way to ensure the smoothness of the transition maps is to take diffeomorphisms as coordinate maps ϕ, ψ. This is done in most cases of everyday practice. In particular, we take diffeomorphisms as coordinate maps throughout this book.

The greater generality gained by allowing the coordinate maps to be homeomorphisms becomes significant when the ideas of coordinate systems and transition maps are applied to more abstract spaces to create the general