Elem. Math. 52 (1997) 45 – 59
0013-6018/97/020045-15 $ 1.50+0.20/0

Elemente der Mathematik

Sattelpunkte oder Variationsprinzipien in Geometrie und Mechanik

Michael Struwe

Michael Struwe wurde 1955 in Wuppertal geboren. Er hat 1980 an der Universität Bonn promoviert. Seit 1986 ist er Professor für Mathematik an der ETH Zürich. Seine Spezialgebiete sind partielle Differentialgleichungen und Variationsrechnung.

1 Einleitung

Der Kreis umschliesst unter allen einfach geschlossenen Kurven in der Ebene zu gegebenem Umfang den grössten Flächeninhalt. Die Bahnen der Planeten folgen dem "Prinzip der kleinsten Wirkung" (Maupertuis); Lichtstrahlen wählen den optisch kürzesten Weg (Fermat). Dem Betrachter scheint es, als wären die Objekte der klassischen Geometrie ebenso wie alle Naturvorgänge durch ihre Optimalität bestimmt; Leibniz folgert kühn, wir lebten in der "besten aller möglichen Welten". Die bodenständige Antwort auf diese These liess in Gestalt von Voltaire's "Candide" nicht lange auf sich warten. Auch aus

Viele Fragen, sei es in Physik, Biologie oder Wirtschaft, lassen sich auf Extremalaufgaben zurückführen, und viele Naturvorgänge lassen sich mit Hilfe von Variationsprinzipien beschreiben. Das "Prinzip der kleinsten Wirkung" der Physik, das isoperimetrische Problem der Geometrie, das "traveling salesman" Problem des Operation Research und Optimierungsvorschläge zur Steigerung der Effizienz öffentlicher Verwaltungen sind nur einige wenige Beispiele, die diese Aussage belegen. Ist in einem solchen Problem eine quantifizierbare Grösse extremal oder allgemeiner stationär zu machen, so sind natürlich mathematische Methoden besonders gefragt. Variationsprobleme, kontinuierliche *und* diskrete, treten aus diesem Grund in der anwendungsorientierten Mathematik ausserordentlich häufig auf. Die Mathematik ist im Laufe der theoretischen Beschäftigung mit diesem Gegenstand zu tiefliegenden Einsichten gelangt, vor allem im kontinuierlichen Fall, und viele der heute verwendeten expliziten Berechnungsmethoden basieren wesentlich auf diesen Erkenntnissen. – In seinem Beitrag geht Michael Struwe auf die theoretische Seite des Gebietes näher ein und zeigt, wie sich hier in reizvoller Weise Analysis, Geometrie und Topologie begegnen. Der Artikel basiert auf Vorträgen des Autors an der Lehrerfortbildungsakademie in Dillingen im November 1995 und im Kolloquium über Mathematik, Informatik und Unterricht an der ETH Zürich im November 1996. *ust*

mathematischer Sicht lässt sich der Anspruch, alle Naturvorgänge liessen sich durch Extremalprinzipien beschreiben, nicht halten. Wie könnten zum Beispiel von einem Punkt ausgehende Lichtstrahlen durch eine Linse in einem zweiten Punkt fokussiert werden und hinter diesem Brennpunkt auseinanderlaufen, wenn alle Lichtstrahlen optisch *kürzeste* Verbindungen aller auf ihnen liegenden Punkte sein sollten?

Dennoch liefern Variationsprinzipien, die Naturphänomene als "kritische Punkte" gewisser Wirkungsfunktionen deuten, eine nahezu umfassende Beschreibung der uns umgebenden Welt. Der Begriff des "kritischen Punktes" muss dazu jedoch weiter gefasst werden.

Betrachten wir als Beispiel die periodische Bewegung eines Massenpunktes in einem konservativen Kraftfeld. In die Sprache der Geometrie übersetzt, handelt es sich um eine geschlossene "Geodäte" auf einer "Energiehyperfläche" im "Phasenraum", der die Ortskoordinaten und die Komponenten des Geschwindigkeitsvektors des Teilchens enthält.

Als Modell für die allgemeine Situation betrachten wir die Sphäre

$$S = \{(x, y, z) \in \mathbb{R}^3; x^2 + y^2 + z^2 = 1\}.$$

Geodäten auf S sind Grosskreisbögen, zum Beispiel Abschnitte der Längenkreise oder des Äquators. Die kürzeste Verbindung zwischen zwei Punkten auf der Sphäre ist stets eine Geodäte – daher führen manche Flugverbindungen von Europa nach Asien über den Nordpol. Jedoch ist nicht jede geodätische Linie auch kürzeste Verbindung ihrer Endpunkte; man gelangt viel schneller von Frankfurt nach Zürich, indem man ein kurzes Stück auf dem gemeinsamen Längenkreis nach Süden fliegt, als durch Wahl des komplementären Bogens auf demselben Längenkreis, welcher über die Pole führt. Insbesondere ist eine geschlossene Geodätische, ein Grosskreis, wo Anfangs- und Endpunkt zusammenfallen, nicht die kürzeste Verbindung zwischen diesen Punkten. Geschlossene Geodäten sind im allgemeinen auch nicht kürzer als jede hinreichend nahe bei ihnen gelegene geschlossene Kurve; zum Beispiel ist der Äquator auf S länger als jeder beliebig nahe beim Äquator gelegene Breitenkreis.

Geschlossene Geodäten auf S sind "Sattelpunkte" der Längenfunktion: Innerhalb einer Schar von Vergleichskurven (den Breitenkreisen) sind sie die längsten, in einer anderen Schar die kürzesten Kurven in der jeweiligen Klasse.

Hier bereits erkennen wir einen fundamentalen Unterschied zu den aus der Schule bekannten Extremwertaufgaben für Funktionen nur einer einzigen reellen Variablen. Während letztere im allgemeinen nur Minima und Maxima als kritische Punkte zulassen, besitzen Funktionen, die von zwei oder mehr Variablen abhängen, zusätzlich kritische Punkte allgemeineren Typs.

Bei den oben erwähnten Problemen haben wir es sogar mit Funktionen zu tun, die von unendlich vielen Variablen abhängen, da man zum Beispiel jeden Punkt einer Bahn auf der Sphäre als eine Variable ansehen kann. In solchen Fällen kann es vorkommen, dass eine Funktion nur Sattelpunkte als kritische Punkte besitzt. Die oben betrachtete Längenfunktion ist zum Beispiel nach oben unbeschränkt, da es zwischen je zwei Punkten auf der Sphäre beliebig lange Verbindungskurven gibt; sie besitzt also kein Maximum. Viele Probleme der klassischen Mechanik werden sogar durch Funktionen beschrieben,

die nach oben *und* unten unbeschränkt sind und daher weder Minimum noch Maximum besitzen.

Sattelpunkte spielen in der Mathematik, insbesondere in der Geometrie, und in deren Anwendungen auf Probleme der Mechanik eine fundamentale Rolle.

Kehren wir zurück zu den Geodäten auf S. Die betrachtete Situation ist aufgrund der Symmetrie der Sphäre hochgradig entartet; jeder Grosskreis auf S ist eine geschlossene Geodäte. Sucht man jedoch einfach geschlossene Geodäten auf dem Ellipsoid

$$S_{abc} = \{(x,y,z) \in \mathbb{R}^3; \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\}$$

mit Halbachsen $0 < a < b < c$, so findet man nur die drei Schnittkurven von S_{abc} mit den Symmetrie-Ebenen $\{x = 0\}, \{y = 0\}$ oder $\{z = 0\}$. Durch Projektion in die Ebene $\{z = 0\}$ gehen diese über in die elliptische Randkurve eines ebenen "Billards" und ein Paar gerader Linien, die diese Randkurve senkrecht treffen und nach dem Stoss an der "Bande" in sich zurücklaufen, vgl. Abbildung 1.

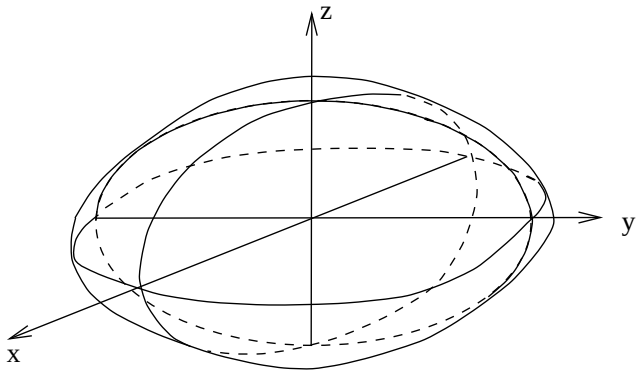

Abb. 1

Wiederum erhalten wir im "entarteten" Fall $a = b = c = 1$ unendlich viele derartiger Linien, nämlich alle Geraden durch den Mittelpunkt des Einheitskreises. Analog zu dieser Situation sprechen wir daher auch im allgemeinen Fall von "Durchmesserlinien".

Die Aufgabe, in einem ebenen Billard alle Durchmesserlinien zu finden, ist also ein Spezialfall der Aufgabe, alle einfach geschlossenen Geodäten auf einer geschlossenen Fläche im Raum zu finden, und verwandt mit dem Problem der Bestimmung periodischer Bahnen mechanischer Systeme. Im Unterschied zu letzteren Aufgaben führt jedoch die Frage nach den Durchmesserlinien eines ebenen Billards auf ein endlich-dimensionales Variationsproblem mit nur zwei unabhängigen Variablen, welches wir mit einfachen Mitteln vollständig analysieren können und welches bereits alle wesentlichen Phänomene der allgemeinen Situation illustriert.

Im folgenden stellen wir zunächst die nötigsten Begriffe bereit und entwickeln Methoden, wie man im Endlichdimensionalen Sattelpunkte charakterisieren kann. Dabei wird

eine tiefgreifende Beziehung offenbar zwischen Anzahl und Art der Sattelpunkte einer Funktion und gewissen "topologischen" Eigenschaften des Definitionsbereichs, die wir ausnutzen, um für unser Modellproblem einen allgemeinen Satz über die Anzahl der Durchmesserlinien in einem ebenen Billard zu folgern.

2 Konzepte und Definitionen

Der Einfachheit halber betrachten wir in diesem Abschnitt nur beliebig oft differenzierbare "glatte" Funktionen $f: \mathbb{R}^2 \to \mathbb{R}$ in zwei reellen Veränderlichen. Analog kann man reelle Funktionen in n Variablen oder Funktionen auf Gebieten, Hyperflächen und Untermannigfaltigkeiten des $\mathbb{R}^n$ behandeln.

Bezeichnen wir die Koordinaten in der euklidischen Ebene wie üblich mit den Symbolen x und y, so können wir die *partiellen Ableitungen* von f in einem Punkt $(x_0, y_0) \in \mathbb{R}^2$ bilden, indem wir die nach der reellen Variablen t differenzierbaren Hilfsfunktionen $t \mapsto \varphi(t) = f(x_0 + t, y_0)$, bzw. $t \mapsto \psi(t) = f(x_0, y_0 + t)$ einführen und setzen

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{d}{dt}\varphi(t)_{|t=0},$$
$$\frac{\partial f}{\partial y}(x_0, y_0) = \frac{d}{dt}\psi(t)_{|t=0}.$$

Der *Gradient* von f im Punkt (x_0, y_0) ist der Vektor

$$\nabla f(x_0, y_0) = \left(\frac{\partial f}{\partial x}(x_0, y_0), \frac{\partial f}{\partial y}(x_0, y_0)\right).$$

Von besonderer Bedeutung für das folgende ist die *Kettenregel*: Für jede glatte Kurve $\gamma: \mathbb{R} \to \mathbb{R}^2$ mit Komponenten $\gamma(t) = (\gamma^1(t), \gamma^2(t))$ ist die zusammengesetzte Funktion $f \circ \gamma$, definiert durch $(f \circ \gamma)(t) = f(\gamma(t))$ für alle t, differenzierbar, und es gilt

$$\frac{d}{dt}(f \circ \gamma)(t) = \nabla f(\gamma(t)) \cdot \frac{d}{dt}\gamma(t). \tag{1}$$

Dabei bezeichnet $a \cdot b = a^1 b^1 + a^2 b^2$ das Skalarprodukt der Vektoren $a = (a^1, a^2), b = (b^1, b^2) \in \mathbb{R}^2$; das heisst, $\frac{d}{dt}(f \circ \gamma)(t)$ lässt sich auch in der Form

$$\frac{d}{dt}(f \circ \gamma)(t) = \frac{\partial f}{\partial x}(\gamma(t))\frac{d}{dt}\gamma^1(t) + \frac{\partial f}{\partial y}(\gamma(t))\frac{d}{dt}\gamma^2(t)$$

schreiben.

Sei $\nabla f(x_0, y_0) \neq 0$. Mittels (1) können wir $\nabla f(x_0, y_0)/|\nabla f(x_0, y_0)|$ als die "Richtung steilsten Anstiegs" der Funktion f deuten: Unter allen Kurven $\gamma: \mathbb{R} \to \mathbb{R}^2$ mit $\gamma(0) = (x_0, y_0)$ und $|\frac{d}{dt}\gamma(0)| = 1$ ist die Zuwachsrate $\frac{d}{dt}(f \circ \gamma)(0)$ am grössten, falls

$$\frac{d}{dt}\gamma(0) = \frac{\nabla f(x_0, y_0)}{|\nabla f(x_0, y_0)|}.$$

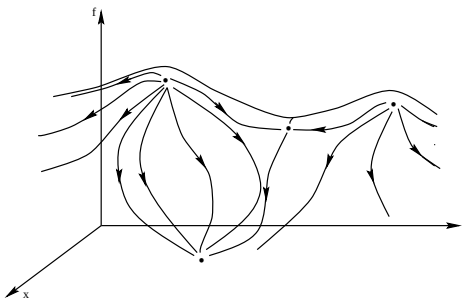

Abb. 2

Besonders anschaulich wird dies, wenn wir uns den Graphen von f als ein Geländerelief

$$\mathcal{G}(f) = \{(x, y, f(x, y)); (x, y) \in \mathbb{R}^2\}$$

im 3-dimensionalen Raum vorstellen, indem wir $f(x, y)$ als die "Höhe" eines Punktes mit ebenen Koordinaten (x, y) deuten, siehe Abbildung 2. Die Punkte $(x_0, y_0) \in \mathbb{R}^2$ mit $\nabla f(x_0, y_0) = 0$ bezeichnen wir als "kritisch". Sie entsprechen in diesem Bild genau den Punkten von $\mathcal{G}(f)$ mit horizontaler Tangentialebene; ein imaginärer (punktförmiger) Wassertropfen verharrt an diesen Punkten in Ruhe, während er von einem Punkt mit $\nabla f(x_0, y_0) \neq 0$ unter Einwirkung der Schwerkraft in Richtung des negativen Gradienten fortfliesst. Das Bild einer Wasserströmung auf dem Relief $\mathcal{G}(f)$ erlaubt nun eine weitere Einteilung der kritischen Punkte von f in "Quellen", von denen das Wasser wegfliesst, "Senken", in denen sich das Wasser sammelt, und "Sattel", an denen sich die Strömung teilt. Mathematisch entsprechen Quellen natürlich den relativen Maxima von f, also den Punkten $z_0 = (x_0, y_0)$, so dass für alle $z = (x, y)$ in der Nähe von z_0 die Beziehung $f(z) \leq f(z_0)$ gilt. Analog entsprechen Senken den relativen Minima z_0 von f mit $f(z) \geq f(z_0)$ für alle z nahe z_0.

Das lokale Verhalten von f in der Nähe eines Sattelpunktes kann im allgemeinen sehr kompliziert sein. "Generisch" verhält sich jedoch eine glatte Funktion f in der Nähe eines Sattelpunktes (gegebenenfalls nach Verschiebung des Nullpunktes, Drehung des Koordinatensystems und Streckung der Koordinatenachsen) wie die Funktion

$$f(x, y) = \frac{1}{2}(x^2 - y^2), \tag{2}$$

deren Richtungsfeld $\nabla f(x, y) = (x, -y)$ auch eine gute Vorstellung vom Strömungsverlauf in der Umgebung eines solchen "nicht entarteten" Sattelpunktes liefert.

Sattelpunkte lassen sich im allgemeinen nicht durch Störung von f beseitigen sondern nur verschieben. Dies kann man auch experimentell leicht verifizieren, indem man das Höhenrelief $\mathcal{G}(f)$ einer Funktion f mit einer Plasticplane modelliert, auf die man Wasser oder Sand "regnen" lässt. Sattelpunkte sind daher wesentlich verschieden von den "Stufenpunkten" reeller Funktionen $f: \mathbb{R} \to \mathbb{R}$, die man durch beliebig kleine Störungen beseitigen oder in ein Paar von relativen Minima und Maxima auflösen kann.

Schliesslich definieren wir noch: $\beta \in \mathbb{R}$ heisst *kritischer Wert* für f, falls es einen kritischen Punkt z_0 von f gibt mit $f(z_0) = \beta$; sonst heisst β ein *regulärer Wert* von f.

3 Kritische Punkte und "Topologie"

Über die Existenz von Minima und Maxima glatter Funktionen gibt der folgende Satz von Weierstrass erschöpfend Auskunft.

Satz: (Weierstrass) *Eine stetige Funktion $f: S \to \mathbb{R}$ auf einer kompakten Menge S besitzt stets ein Minimum und ein Maximum.*

Beispiele für kompakte Flächen im $\mathbb{R}^3$ sind die Sphäre, der Torus oder der g-Torus mit g Löchern, $g \geq 2$. "Legen" wir diese Flächen geeignet in den 3-dimensionalen Raum und betrachten die z-Koordinate eines Punktes als unsere Funktion $f: S \to \mathbb{R}$, so sehen wir, dass eine Funktion auf der Sphäre im allgemeinen nur zwei kritische Punkte (mit horizontaler Tangentialebene an S) besitzt, und zwar die vom Satz von Weierstrass geforderten Minima und Maxima. Eine Funktion auf dem Torus hat jedoch im allgemeinen, scheint es, noch zusätzlich zwei Sattelpunkte, und für jedes weitere Loch, welches S aufweist, kommt ein weiteres Paar von Sattelpunkten hinzu; siehe Abbildung 3.

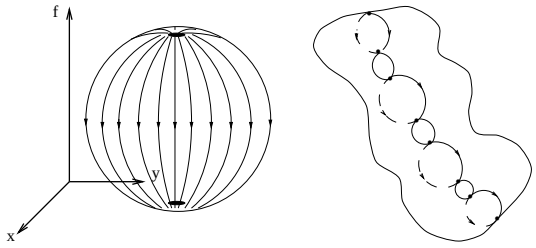

Abb. 3

Gibt es also eine Beziehung zwischen der "Topologie" einer Menge (welche im vorliegenden Beispiel bequem durch die Zahl der Löcher charakterisiert werden kann) und der Zahl der Sattelpunkte, die *jede* auf dieser Menge definierte Funktion (mindestens) haben muss? Und schliesslich: Gibt es ein systematisches Verfahren, um diese Sattelpunkte, wenn es sie gibt, zu finden? – Diesen Fragen wollen wir in den nächsten Abschnitten nachgehen. Damit das Vorhaben unsere Mittel nicht übersteigt, beschränken wir uns dabei im wesentlichen auf den Torus, den wir auch erhalten können, indem wir auf dem Einheitsquadrat $Q = \{(x, y) \in \mathbb{R}^2; 0 \leq x, y \leq 1\}$ gegenüberliegende Punkte auf dem Rande identifizieren, vgl. Abbildung 4.

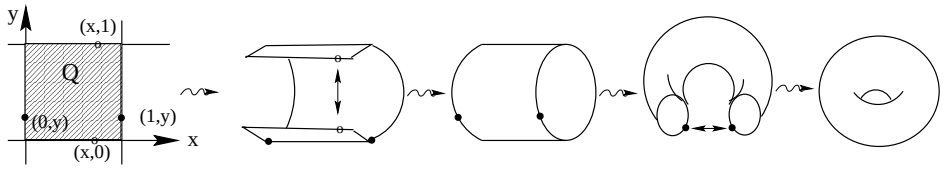

Abb. 4

Beachten wir nun noch, dass eine in beiden Variablen mit der Periode 1 periodische Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ durch ihre Werte auf Q vollkommen bestimmt ist, und dass wir umgekehrt jede Funktion $f\colon Q \to \mathbb{R}$ mit $f(0,y) = f(1,y)$ sowie $f(x,0) = f(x,1)$ für $0 \le x, y \le 1$ periodisch zu einer doppelt periodischen Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ fortsetzen können, so entspricht das Studium reeller Funktionen auf dem Torus der Untersuchung doppelt periodischer Funktionen $f\colon \mathbb{R}^2 \to \mathbb{R}$ mit

$$f(x+k, y+l) = f(x,y) \text{ für } k,l \in \mathbb{Z};$$

wir schreiben hierfür auch $f\colon \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}$. Später werden wir sehen, dass diese Periodizitätsbedingung im eingangs formulierten Modellproblem erfüllt ist.

4 Der Gradientenfluss

Einen systematischen Zugang zum Problem, alle kritischen Punkte, insbesondere die Sattelpunkte einer gegebenen Funktion f zu finden, erhalten wir, indem wir unser obiges Bild von einer dem Gefälle folgenden Strömung auf dem durch den Graphen von f gegebenen Höhenrelief formalisieren.

Der Einfachheit halber beschränken wir uns auf doppelt periodische glatte Funktionen $f\colon \mathbb{R}^2 \to \mathbb{R}$. Der Gradient von f definiert das (ebenfalls doppelt periodische) Richtungsfeld

$$e = -\nabla f\colon \mathbb{R}^2 \to \mathbb{R}^2.$$

Zu gegebenem Anfangspunkt $z_0 = (x_0, y_0)$ bestimmen wir die *Integralkurve* $\gamma = \gamma(\cdot; z_0)\colon \mathbb{R} \to \mathbb{R}^2$ von e durch z_0 als Lösung des Anfangswertproblems

$$\frac{d}{dt}\gamma(t) = e(\gamma(t)), \tag{3}$$

$$\gamma(0) = z_0. \tag{4}$$

Die Existenz und Eindeutigkeit von γ folgt aus allgemeinen Sätzen über gewöhnliche Differentialgleichungen. Insbesondere gilt aufgrund der Eindeutigkeit der Lösung von (3), (4) für alle $z_0 = (x_0, y_0) \in \mathbb{R}^2$ und $s, t \in \mathbb{R}$ die Beziehung

$$\gamma(s; \gamma(t; z_0)) = \gamma(s+t; z_0); \tag{5}$$

das heisst, die Bahn durch den Punkt $\gamma(t; z_0)$ ist die um t zeitverschobene Fortsetzung der Bahn durch den Punkt z_0. Weiter gilt aufgrund der Periodizität von e für alle $z_0 = (x_0, y_0) \in \mathbb{R}^2$ und alle $(k,l) \in \mathbb{Z}^2$

$$\gamma(t; (x_0 + k, y_0 + l)) = \gamma(t; (x_0, y_0)) + (k,l). \tag{6}$$

Um ein qualitatives Bild vom Verlauf dieser Bahnen in der Nähe eines kritischen Punktes zu erhalten, betrachten wir als Beispiel die Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$, gegeben durch

$$f(x,y) = \frac{1}{2}(\alpha x^2 + \beta y^2), (x,y) \in \mathbb{R}^2,$$

mit Parametern $\alpha, \beta \in \mathbb{R}$.

Das durch f definierte Richtungsfeld $e\colon \mathbb{R}^2 \to \mathbb{R}^2$ hat die Darstellung

$$e(x,y) = -\nabla f(x,y) = (-\alpha x, -\beta y).$$

Bezeichnen wir die Komponenten der durch (3), (4) bestimmten Kurve γ durch einen Punkt $z_0 = (x_0, y_0)$ mit $\gamma(t) = (x(t), y(t))$, so geht (3), (4) über in das Paar gewöhnlicher Differentialgleichungen

$$\frac{d}{dt}x = -\alpha x,\ x(0) = x_0; \quad \frac{d}{dt}y = -\beta y,\ y(0) = y_0.$$

Als Lösung erhalten wir

$$x(t) = x_0 e^{-\alpha t},\ y(t) = y_0 e^{-\beta t}.$$

Speziell für die Fälle $\alpha = \beta = 1,\ \alpha = \beta = -1$ und $\alpha = 1, \beta = -1$ ergeben sich somit die typischen Bilder des Strömungsverlaufs in der Nähe einer Senke (Minimum), einer Quelle (Maximum), beziehungsweise eines Sattels; siehe Abbildung 5.

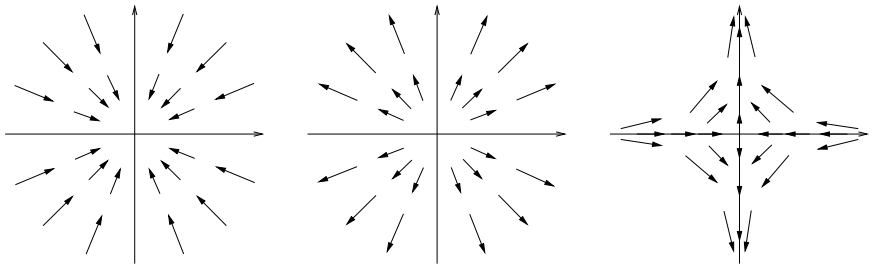

Abb. 5

Jede Lösung $\gamma = \gamma(\cdot; z_0)$ von (3), (4) lässt sich deuten als Stromlinie eines Teilchens, das sich zur Zeit $t = 0$ an der Stelle z_0 befindet; andererseits definiert die Gesamtheit aller dieser Bahnen eine vom Parameter t abhängige Schar von Abbildungen $\Phi(\cdot, t)\colon \mathbb{R}^2 \to \mathbb{R}^2$, indem wir für jedes $t \in \mathbb{R}$ definieren

$$\Phi(z_0, t) = \gamma(t; z_0), z_0 \in \mathbb{R}^2.$$

Da $\Phi(\cdot, t)$ aufgrund von (6) für festes t mit Verschiebungen um $(k,l) \in \mathbb{Z}^2$ vertauscht, können wir $\Phi(\cdot, t)$ auch als Abbildung $\Phi(\cdot, t)\colon \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ auffassen, das heisst, als eine Transformation des Torus.

Sehen wir nun sowohl z als auch t als variabel an, so erhalten wir eine Abbildung $\Phi\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$, den *Gradientenfluss* zu f, welcher sowohl die Schar von Abbildungen $(\Phi(\cdot, t))_{t\in\mathbb{R}}$ als auch die individuellen Stromlinien $\gamma(\cdot; z_0) = \Phi(z_0, \cdot)$ durch jeden Punkt $z_0 \in \mathbb{R}^2$ erzeugt. Da wir f als glatt voraussetzen, ist auch Φ eine glatte Funktion in allen Variablen.

Weiter können wir (3), (4) äquivalent ausdrücken durch die Bedingung

$$\frac{\partial}{\partial t}\Phi = e \circ \Phi,\ \Phi(\cdot, 0) = id. \tag{7}$$

Schliesslich besitzt die Schar $(\Phi(\cdot, t))_t$ aufgrund von (5) die Eigenschaft

$$\Phi(\cdot, s) \circ \Phi(\cdot, t) = \Phi(\cdot, s + t). \tag{8}$$

Insbesondere ist jede Abbildung $\Phi(\cdot, t)$ stetig invertierbar mit $\Phi(\cdot, t)^{-1} = \Phi(\cdot, -t)$; das heisst, $\Phi(\cdot, t)$ ist ein Homöomorphismus.

Analog erhält man für glatte Funktionen f auf einer Fläche S einen Gradientenfluss $\Phi: S \times \mathbb{R} \to S$ mit den obigen Eigenschaften.

5 Minimax-Prinzip

Mit Hilfe des Gradientenflusses $\Phi: S \times \mathbb{R} \to S$ zu $f: S \to \mathbb{R}$ kann man nicht nur einzelne Punkte sondern auch Teilmengen von S "transportieren". Geeignete Mengen bleiben dabei an Sattelpunkten "hängen".

Um dies zu veranschaulichen, betrachten wir als Beispiel für S einen in $\mathbb{R}^3$ eingebetteten Torus, f die z-Koordinate eines Punktes. Sei $A_0 = \alpha_0(\mathbb{R})$ Bild einer geschlossenen Kurve $\alpha_0: \mathbb{R} \to S$ auf S, die sich auf S nicht in einen Punkt "zusammenziehen" lässt.

Dann erhalten wir eine Schar derartiger Mengen, indem wir setzen

$$A_t = \Phi(A_0, t) = \alpha_t(\mathbb{R}), \alpha_t = \Phi(\cdot, t) \circ \alpha_0.$$

Wir erwarten, dass für $t \to \infty$ die Schar A_t gegen eine Grenzkurve A_∞ strebt, die sich im tiefsten Sattelpunkt um den Torus schlingt; siehe Abbildung 6.

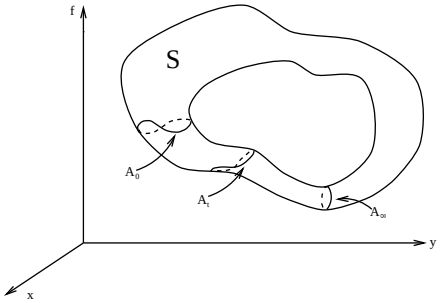

Abb. 6

Allgemeiner definieren wir eine Familie $\mathcal{A}$ von Teilmengen $A \subset S$ als (vorwärts-) Φ-*invariant*, falls $\Phi(A, t) \in \mathcal{A}$ für alle $A \in \mathcal{A}$ und $t \geq 0$.

Offenbar ist $\mathcal{A} = \{S\}$ eine Φ-invariante Familie, ebenso $\mathcal{A} = \{\{z\}; z \in S\}$. Jede Menge $A = A_0$ erzeugt wegen (8) eine Φ-invariante Familie

$$\mathcal{A} = \{\Phi(A, t); t \geq 0\}.$$

Nun können wir das Hauptresultat dieses Abschnitts formulieren.

Minimax-Prinzip. *Sei $f: S \to R$ eine glatte Funktion auf der kompakten Fläche S, Φ der Gradientenfluss zu f, und sei $\mathcal{A}$ eine Φ-invariante Familie in S.*

Dann ist

$$\beta = \inf_{A \in \mathcal{A}} \sup_{a \in A} f(a)$$

ein kritischer Wert von f.

Zum Beispiel erhalten wir im Falle $\mathcal{A} = \{S\}$ den kritischen Wert

$$\overline{\beta} = \max_{a \in S} f(a);$$

im Falle $\mathcal{A} = \{\{a\}; a \in S\}$ hingegen

$$\underline{\beta} = \min_{a \in S} f(a).$$

Falls S der Torus ist, erwarten wir zudem, mit der am Anfang dieses Kapitels beschriebenen Konstruktion auch kritische Werte $\underline{\beta} < \beta < \overline{\beta}$ zu erhalten, welche Sattelpunkten entsprechen.

Zuvor wollen wir jedoch zumindest für den Fall des Torus einen Beweis des Minimaxprinzips angeben. Sei also $f: \mathbb{R}^2 \to \mathbb{R}$ doppelt periodisch, $\mathcal{A}$ eine Φ-invariante Familie. Wir argumentieren indirekt und nehmen widerspruchsweise an, β sei regulär.

Es folgt, dass für eine geeignete Zahl $\varepsilon > 0$ und alle $z \in \mathbb{R}^2$ gilt:

$$|f(z) - \beta| < \varepsilon \Rightarrow |\nabla f(z)|^2 > \varepsilon. \tag{9}$$

Andernfalls gäbe es eine Folge von Punkten $z_n \in \mathbb{R}^2$ mit

$$f(z_n) \to \beta, \nabla f(z_n) \to 0 (n \to \infty). \tag{10}$$

Wegen der Periodizität von f dürfen wir annehmen, dass die Folge (z_n) beschränkt ist. Aufgrund des Satzes von Weierstrass besitzt (z_n) dann einen Häufungspunkt z, und es gibt eine Teilfolge $(z_n)_{n \in \Lambda}$ mit $z_n \to z (n \to \infty, n \in \Lambda)$. Grenzübergang in (10) für $n \to \infty, n \in \Lambda$ liefert dann wegen der Stetigkeit von f und ∇f die Gleichung $f(z) = \beta, \nabla f(z) = 0$; das heisst, β ist kritisch, im Widerspruch zu unserer Annahme. Damit ist (9) gezeigt.

Für $A \in \mathcal{A}$ mit

$$\sup_{a \in A} f(a) < \beta + \varepsilon \tag{11}$$

und alle $a \in A$ mit

$$f(a) > \beta - \varepsilon \tag{12}$$

folgt nun aus (1), (7) und (9)

$$\frac{d}{dt} f(\Phi(a,t))_{|t=0} = \nabla f(a) \cdot \frac{\partial}{\partial t} \Phi(a,t)_{|t=0} = -|\nabla f(a)|^2 < -\varepsilon. \tag{13}$$

Insbesondere gilt (11) auch für alle Mengen $A_t = \Phi(A,t), t \geq 0$.

Beachte nun, dass $A_t \in \mathcal{A}$ für alle $t \geq 0$, da $\mathcal{A}$ nach Voraussetzung Φ-invariant ist. Somit gilt nach Definition von β auch stets

$$\beta(t) := \sup_{a \in A_t} f(a) \geq \beta,$$

und zur Bestimmung von $\beta(t)$ muss man nur Punkte $a \in A_t$ berücksichtigen, die auch (12) und daher (13) erfüllen.

Es folgt, die Funktion $t \mapsto \beta(t)$ ist monoton fallend, und

$$\frac{d}{dt}\beta(t) \leq -\varepsilon.$$

Nach der Zeit $t = 1$ erhalten wir den Widerspruch

$$\beta \leq \beta(1) = \sup_{a \in A_1} f(a) \leq \beta(0) - \varepsilon = \sup_{a \in A} f(a) - \varepsilon < \beta. \qquad \square$$

Die Anwendung des Minimaxprinzips für reelle Funktionen auf dem Torus wollen wir nun anhand unseres Modellproblems illustrieren.

6 Anwendung auf Modellproblem

Sei Γ die Randkurve eines konvexen ebenen Billards, parametrisiert durch eine glatte Abbildung $\gamma\colon \mathbb{R} \to \mathbb{R}^2$ mit $\gamma(x+1) = \gamma(x)$ für alle x. Weiter nehmen wir an, dass γ nach Bogenlänge parametrisiert ist, das heisst, $|\gamma'(x)| = 1$ für alle x, und dass γ im Intervall $[0, 1[$ injektiv ist, also keine Doppelpunkte besitzt. Wir identifizieren "Bahnen" zwischen Punkten $p = \gamma(x)$ und $q = \gamma(y)$ auf Γ mit dem Paar (x, y). Aufgrund der Periodizität von γ liefern Paare (x, y), bzw. $(x', y') \in \mathbb{R}^2$ dieselbe Bahn, falls $(x - x', y - y') \in \mathbb{Z}^2$; vgl. Abbildung 7.

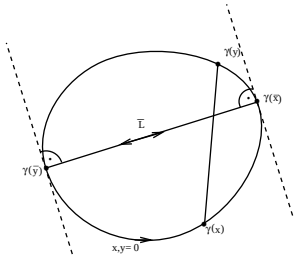

Abb. 7

Offenbar geht eine Bahn (x, y) nach Reflexion in Γ in sich über, falls die Verbindungsgerade von $\gamma(x)$ nach $\gamma(y)$ den Rand senkrecht trifft, das heisst, falls gilt

$$(\gamma(y) - \gamma(x)) \cdot \gamma'(x) = (\gamma(y) - \gamma(x)) \cdot \gamma'(y) = 0. \qquad (14)$$

Sei nun $f: \mathbb{R}^2 \to \mathbb{R}$ definiert durch

$$f(x,y) = \frac{1}{2}|\gamma(x) - \gamma(y)|^2.$$

Aufgrund der Periodizität von γ ist f doppelt periodisch, $f: \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}$. Zusätzlich weist f die folgende Symmetrie auf

$$f(x,y) = f(y,x). \tag{15}$$

Weiter gilt:

$$\frac{\partial f}{\partial x}(x,y) = (\gamma(x) - \gamma(y)) \cdot \gamma'(x),$$
$$\frac{\partial f}{\partial y}(x,y) = -(\gamma(x) - \gamma(y)) \cdot \gamma'(y);$$

das heisst, Durchmesserlinien gehören zu kritischen Punkten von f.

Betrachten wir zunächst die Maxima und Minima von f. Offenbar gilt $f(x,y) \geq 0$ für alle (x,y), und $f(x,y) = 0$ genau dann, wenn $x - y \in \mathbb{Z}$. Die Minima von f entsprechen also genau den konstanten "Bahnen", wo die Kugel an einem Randpunkt des Billards liegenbleibt. Diese "Bahnen" sind natürlich für uns nicht von Interesse.

Der Satz von Weierstrass liefert uns hingegen auch ein Paar $(\overline{x}, \overline{y}) = \overline{z}$ mit

$$f(\overline{z}) = \max_{z \in \mathbb{R}^2} f(z) = \overline{\beta}.$$

Geometrisch entspricht $(\overline{x}, \overline{y})$ der längsten Durchmesserlinie $\overline{L}$ oder der Richtung, in welcher Γ die maximale "Dicke" aufweist. Aufgrund der Periodizität von f finden wir unendlich viele weitere Maxima, die aus $(\overline{x}, \overline{y})$ oder $(\overline{y}, \overline{x})$ durch Translation mit einem Paar $(k,l) \in \mathbb{Z}^2$ hervorgehen; diese entsprechen jedoch alle derselben Durchmesserlinie. Finden wir mit Hilfe des Minimaxprinzips hiervon auch geometrisch verschiedene weitere kritische Punkte von f?

Um das Minimax-Prinzip einzusetzen, müssen wir eine geeignete Familie von Mengen finden, die unter dem Gradientenfluss invariant ist. Zum Beispiel können wir als $\mathcal{A}$ die Familie der Mengen $A = \alpha([0,1])$ definieren, wobei $\alpha: [0,1] \to \mathbb{R}^2$ ein stetiger "Weg" ist, der zwei fest gewählte Minima $\alpha(0) = (x_0, x_0) = z_0, \alpha(1) = (x_1, x_1 + 1) = z_1$ in verschiedenen "Zusammenhangskomponenten" der Menge $\{(x, x+k); x \in \mathbb{R}, k \in \mathbb{Z}\}$ der Minima von f miteinander verbindet; vgl. Abbildung 8.

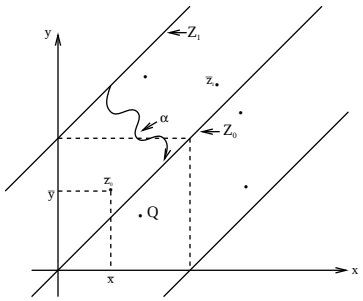

Abb. 8

Offenbar gilt $\Phi(z_0,t)=z_0, \Phi(z_1,t)=z_1$ für alle t. Daher ist mit α auch $\Phi(\cdot,t)\circ\alpha$ ein Weg von z_0 nach z_1, und die so definierte Familie ist Φ-invariant. (Anstelle der Bedingung $\alpha(0)=z_0, \alpha(1)=z_1$ mit festen Punkten z_0, z_1 könnten wir auch lediglich verlangen, dass $\alpha(0)=(x,x), \alpha(1)=(y,y+1)$ für irgendwelche $x,y\in\mathbb{R}$; auch dies liefert eine Φ-invariante Familie von Mengen $A=\alpha([0,1])$, welche die Gerade $Z_0=\{(x,x); x\in\mathbb{R}\}$ mit der Geraden $Z_1=\{(x,x+1); x\in\mathbb{R}\}$ verbinden.)

Auf jedem derartigen Weg muss ein "Wall" der Mindesthöhe

$$\beta=\inf_{A\in\mathcal{A}}\sup_{a\in A} f(a)\geq \inf_{x\in\mathbb{R}} f(x,x+\frac{1}{2})>0$$

überschritten werden, und aufgrund des Minimax-Prinzips ist β ein kritischer Wert. Im allgemeinen ist der so gefundene Wert β strikt kleiner als $\overline{\beta}$, entspricht also einer Durchmesserlinie L, welche kürzer ist als $\overline{L}$ und damit von $\overline{L}$ verschieden.

Was kann man jedoch aussagen, falls $\beta=\overline{\beta}=\max f$? In diesem Fall trifft *jeder* Weg von z_0 nach z_1 auf ein Maximum von f; insbesondere lassen sich dann zwei Maxima $\overline{z}_0=(\overline{x},\overline{y})$ und $\overline{z}_1=(\overline{x}+1,\overline{y}+1)$ in der Menge der Maxima "verbinden". Geometrisch bedeutet dies, es gibt Durchmesserlinien in jeder Richtung, welche alle dieselbe (maximale) Länge besitzen. (Jedoch folgt hieraus nicht, dass Γ ein Kreis ist. Der Wankelmotor zum Beispiel benützt zu seiner Funktion, dass auch eine aus drei Kreisbögen von je 60° zusammengefügte Kurve konstante Dicke aufweist.)

In jedem Fall zeigt unser Argument jedoch, dass eine ebene Kurve stets (mindestens) zwei Durchmesserlinien besitzt.

7 Topologische Betrachtungen

In Abschnitt 3 haben wir vermutet, es könne eine Beziehung geben zwischen der Zahl der Sattelpunkte einer Funktion $f: S\to\mathbb{R}$ und der "Topologie" von S. Im Falle des ebenen Billards ist nun bereits das Niveau des Minimums $f=0$ hochgradig entartet, und es gibt unendlich viele "triviale" kritische Punkte. Jedoch wollen wir hier versuchen, eine topologische Beziehung aufzudecken, die uns im Fall des ebenen Billards ein Paar von *nicht* trivialen kritischen Punkten liefert. Wesentlich ist die folgende Beobachtung.

Die in Abschnitt 6 untersuchte Funktion $f:\mathbb{R}^2\to\mathbb{R}$ weist neben der doppelten Periodizität auch noch die Spiegelinvarianz

$$f(x,y)=f(y,x)$$

als weitere Symmetrie auf. Daher ist f bereits durch die Werte auf dem Bereich $D=\{(x,y)\in\mathbb{R}^2; 0\leq x\leq y\leq 1\}$ vollkommen bestimmt, dessen Randpunkte zudem durch f folgendermassen "identifiziert" werden

$$f(x,x)=0,\ f(0,y)=f(y,0)=f(y,1),\ 0\leq x,y\leq 1.$$

Kollabieren wir die Punkte auf der Diagonalen $Z_0=\{x=y\}$ in einen einzigen Punkt p_0, so geht der Bereich D über in einen Kreis, wobei f in einander gegenüberliegenden

"Antipoden" auf dem Rand dieselben Werte annimmt. Nun verformen wir den Kreis zu einer Halbsphäre S^+ und vervollständigen S^+ durch "Ankleben" einer zweiten Sphärenkappe zur Sphäre $S \subset \mathbb{R}^3$, wobei wir f durch die Festsetzung $f(p) = f(-p)$ auf S fortsetzen. Nach dieser Operation erhalten wir eine Abbildung $f: S \to \mathbb{R}$, die in *jedem* Paar von Antipoden denselben Wert annimmt und nur in einem einzigen Paar von Punkten $(p_0, -p_0)$ verschwindet. Da Paare von Antipoden $(p, -p)$ auf S genau den Geraden durch den Koordinatenursprung in $\mathbb{R}^3$ entsprechen, deren Gesamtheit den "projektiven Raum" $P^2\mathbb{R}$ bildet, erkennen wir auf diesem Wege, dass unser Modellproblem "topologisch" zu dem Problem äquivalent ist, kritische Punkte von Funktionen $\tilde{f}: P^2\mathbb{R} \to \mathbb{R}$ zu finden; vgl. Abbildung 9.

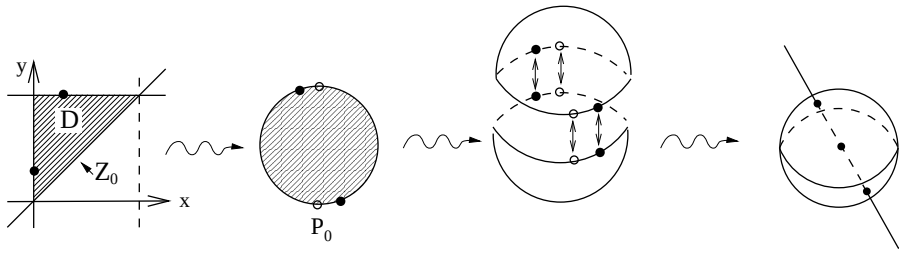

Abb. 9

Der Grund dafür, dass solch eine Funktion ausser dem stets vorhandenen Minimum und Maximum noch mindestens einen weiteren kritischen Punkt besitzt, wird am ehesten ersichtlich, wenn wir zum Bild einer Funktion $f: S \to \mathbb{R}$ zurückkehren, welche die Symmetriebedingung $f(p) = f(-p)$ für alle $p \in S$ erfüllt.

Der von f erzeugte Gradientenfluss Φ weist eine dazu passende Symmetrie auf, und zwar gilt

$$\Phi(p, t) = -\Phi(-p, t)$$

für alle $p \in S, t \in \mathbb{R}$. Folglich erzeugt eine Kurve $\alpha: \mathbb{R} \to S$ mit

$$\alpha(s + 1/2) = -\alpha(s) \tag{16}$$

eine Schar $\alpha_t = \Phi(\cdot, t) \circ \alpha$ von Kurven, die dieselbe Symmetrie aufweisen, und die zugehörige Familie $\mathcal{A} = \{\alpha_t(\mathbb{R}); t \geq 0\}$ ist Φ-invariant. Offenbar verhindert die Bedingung (16), dass sich die Kurven α_t zusammenziehen, und mit dem Minimaxprinzip erhalten wir einen weiteren kritischen Wert. Fällt dieser Wert mit dem Minimum oder dem Maximum von f zusammen, so gibt es unendlich viele Minima, bzw. Maximalstellen von f. Im Falle unseres Modellproblems haben wir alle Minima zu einem einzigen Antipodenpaar kollabiert; daher kann es in diesem Fall nur unendlich viele Maxima, also nicht triviale Lösungen, geben.

Kurven $\alpha: \mathbb{R} \to S$, die der Bedingung (16) genügen, entsprechen übrigens genau den nicht zusammenziehbaren Kurven auf $P^2\mathbb{R}$.

8 Verallgemeinerungen, Ausblick

Mit ähnlichen Konzepten wie den hier vorgestellten beweist man, dass es auf jeder Fläche in $\mathbb{R}^3$ vom Typ der Sphäre (mindestens) drei einfach geschlossene Geodätische gibt, auf jeder topologischen 3-Sphäre mindestens vier "minimale 2-Sphären", usw. Dabei wird man jedoch auf Variationsprobleme in Funktionenräumen geführt, zu deren Behandlung es recht aufwendiger analytischer Methoden bedarf. Variationsmethoden sind auch bei aktuellen Fragestellungen in der Geometrie und in der Theorie der dynamischen Systeme, speziell der Himmelsmechanik, ein wichtiges Hilfsmittel.

Im anschliessenden Literaturverzeichnis sind einige einführende Werke aufgeführt, die einen Überblick über das Gebiet der Variationsrechnung vermitteln und deren Bibliographien eine Übersicht über die umfangreiche Forschungsliteratur zu diesem Thema bieten.

Literatur

[1] S. Hildebrandt und A. Tromba: *Mathematics and optimal form*. Scientific American Books, Inc. New York (1985); Deutsche Übersetzung: Panoptimum, Spektrum der Wissenschaft, Heidelberg (1987).
Neu: *Kugel, Kreis und Seifenblasen. Optimale Formen in Geometrie und Natur*. Birkhäuser, Basel (1996).

[2] J. Mawhin und M. Willem: *Critical point theory and Hamiltonian systems*. Appl. Math. Sci. **74**, Springer, New York-Berlin-Heidelberg-London-Paris-Tokyo (1989)

[3] Struwe, M.: *Variational methods*, 2. Auflage, Ergebnisse der Mathematik und ihrer Grenzgebiete **34** (1996), Springer, Berlin, etc.

Michael Struwe
Mathematik
ETH-Zentrum
CH-8092 Zürich