Introductory note to 1902d

Rüdiger Thiele

In 1899, Zermelo completed his *Habilitation* in Göttingen, becoming there a *Privatdozent* with the right to give lectures. Zermelo had also lectured on the calculus of variations a number of times at the Göttingen Mathematical Society, beginning in 1897, and also, in particular, on Kneser's article "Zur Variationsrechnung" ("On the calculus of variations") (*Kneser 1898*), an important related work. Zermelo's first lectures in Göttingen on the calculus of variations in the summer semester of 1902 and his paper "Zur Theorie der kürzesten Linien" ("On the theory of shortest paths¨") (*Zermelo 1902d*) should be viewed in relationship to the Society lectures; Zermelo's next lecture on the calculus of variations took place in the winter semester of 1907.

The problem of shortest paths or, as one has said since Joseph Liouville (1844), the problem of geodesics on a surface, was identified by Zermelo as one of the classical problems in the calculus of variations, and indeed, questions relating to intuitively geometric geodesics belong among its standard problems. The variational problem leading to geodesics for a parametrically given shortest curve C_0: $(x^0(t), y^0(t))$, $t \in [t_1, t_2]$, on a surface F between the points P_1 and P_2 is as follows:

$$\begin{aligned} J(C) &= \int_C \sqrt{dx^2 + dy^2}\, dt \\ &= \int_{t_1}^{t_2} \sqrt{E\dot{x}^2 + 2F\dot{x}\dot{y} + G\dot{y}^2}\, dt\,, \end{aligned} \tag{1}$$

where E, F, and G in the second formulation of the problem are the Gaussian fundamental coefficients of the surface F.[1]

Let us begin by briefly considering how the problem of shortest paths was treated in Göttingen. In his Paris address "Mathematische Probleme" (1900), David Hilbert, in his Problem 4 ("Problem of the line as the shortest distance between two points"),[2] called "attention to a theorem that is given by many authors as the definition of a straight line, which asserts that the line is the shortest distance between two points" (*1900a*, 268). Hilbert's investigations of this minimality condition are closely related to his work on the foundations of

[1] Zermelo discusses the extension of the variational problem for shortest paths from geometric viewpoints, which is only natural. We therefore formulate the variational problem parametrically.

[2] Along with thirteen further problems, this problem was not presented in the Paris lecture, but appeared only in the printed version of the lecture (*Hilbert 1900a*).

H.-D. Ebbinghaus, A. Kanamori (Eds.), *Ernst Zermelo - Collected Works/ Gesammelte Werke II*, DOI 10.1007/978-3-540-70856-8_8, Schriften der Mathematisch-naturwissenschaftlichen Klasse der Heidelberger Akademie der Wissenschaften 23,

geometry (*1899*). As a result of this work, he developed, among other things, a geometry in which all the axioms of Hermann Minkowski's pseudogeometry are satisfied except for the parallel postulate (*1895*, 91). Hilbert remarked, "In the case of the plane and assuming the axiom of continuity, this problem leads to the question treated by Darboux (*1894*, 54) of finding all variational problems in the plane whose solutions are all the straight lines in the plane."

In his Göttingen dissertation *Über die Geometrien, in denen die Geraden die Kürzesten sind* (*On geometries in which straight lines are shortest paths*) (*Hamel 1901*), written under the supervision of Hilbert, which appeared two years later as *Hamel 1903*, Georg Hamel studied Hilbert's problem from the viewpoint of the calculus of variations. He thus used methods whose analytic assumptions are unsuited to geometry. To be sure, measuring lengths is a very elementary process, but the functional relationships among measurements pertain to arithmetic, not geometric, relationships; the length of a curve can be defined by an integral, as in (1), or even more generally. Using direct methods, Hilbert showed (*1900b*) under weakened assumptions that a given continuous curve on a surface $z = f(x, y)$ that joins two distinct points of the surface is rectifiable; f is taken to be continuous and to have continuous partial derivatives of first order.

The broadly posed fourth problem was made more precise in numerous papers, including those by Paul Funk (*1929*) and Herbert Busemann (*1943*). The desire to formulate assertions of the calculus of variations geometrically was realized, for example, in Finsler geometry, which originated in Paul Finsler's Göttingen dissertation *Über Kurven und Flächen* (*On curves and surfaces*) (*1918*) written under the supervision of Constantin Carathéodory.

Zermelo mentions three possible ways to extend the variational problem $J(C) \to$ minimum associated with (1), which can be characterized by the following key words:

(a) absolute minima,
(b) restrictions on surfaces,
(c) differential inequalities as constraints.

The last two cases appear naturally in practical questions, and in this case, Zermelo introduced the problem of road and rail construction. The construction of roads and rail lines can be limited by geographical conditions to a surface patch $F_0 \subseteq F$; moreover, the steepness or curvature of a road or track section can be constrained. Since Zermelo assumes simply connected surface patches F_0, obstacles on F_0 are excluded from consideration. The constraints (b) and (c), Zermelo observes, can also lead to the result that on the surface F itself, no solutions are possible (and from an engineering perspective, this leads to tunnels and viaducts).[3] However, in the case of constraints, it would be necessary to deal with the question whether under such conditions, a set

[3] About 40 years later, Zermelo would choose this topic as one of the chapters of a planned book *Mathematische Miniaturen* (*Mathematical miniatures*) under

of admissible comparison functions (comparison curves) is available to make possible the variation and its infinitesimal techniques. The standard methods of the calculus of variations require, in contrast to the theory of optimization, that open sets be used. In the case of domain restriction, where also parts of the boundary ∂F_0 of the surface patch F_0 may need to be considered, some special ideas are necessary, which Zermelo sketches for simply connected surface patches F_0 (see also *Bolza 1909*, 392–407, 527).

Zermelo calls curves of constant steepness *Kletterkurven* (*climbing curves*). At every point of the surface there are two simply infinite families of climbing curves, where by alternate use of these curves from both families, a zigzag line can be created with whose help one can approximate curves with larger steepness (see also *Bolza 1909*, 126f.).

A functional $J(x)$ defined on a set A of functions (or curves) can be associated with a variational problem

$$J(x) \to \text{ extremum} \qquad \text{on } A_0 \subseteq A.$$

Here the subset A_0, with which also a notion of neighborhood is associated, determines the type of solution, that is, whether one has a strong or weak, relative or absolute, and so forth, extremum (see *du Bois-Reymond 1879a*, 283; *Kneser 1900*, §17; *Osgood 1900/01*, 105). In his dissertation, Zermelo worked with very general notions of distance and in this way dealt with the question of the type of an extremum in variational problems (*Zermelo 1894*, 24). Of the notions of distance discussed by Zermelo, today only the distances in the space of continuous or continuously differentiable functions are used, which lead to strong and weak extrema.

In the present work, Zermelo looks at relative and absolute minima of geodesics, which he calls *shortest* and *by far shortest* paths, respectively. The infinitesimal variational technique first takes into account only sufficiently close comparison functions (respectively comparison curves) that lie in a given neighborhood of the extremal E (geometrically in a strip around the extremal E), leading thereby first to necessary conditions for relative extrema, in this case minima. Here it is necessary in addition to include non-analytic

the title "Straßenbau im Gebirge" ("Road construction in the mountains"); cf. *Ebbinghaus 2007*, 14.

functions. The extremals E obtained in relation to the strip, however, even if they exhibit minimality, need not lead to (relative) minima for the entire function (or along the entire curve); rather, the presence of minimality can definitely be limited to parts of the function (respectively the curve). In the latter case, the minimality on E is limited by a point P_1^*, called the conjugate to the initial point P_1 of the extremal E. (One would have respectively $P_1^* = (x(t_1^*), y(t_1^*))$ and $P_1^* = (x_1^*, y(x_1^*))$ for parametrically and non-parametrically formulated variational problems.) Conjugate points can be computed from nontrivial solutions $u(t)$ of a Jacobian accessory differential equation if one determines t_1^* with $u(t_1^*) = 0$. The so-called Jacobi differential equation follows from consideration of the second variation (see *Bolza 1909*, §29, for parametric problems, and §12 for non-parametric problems; *Bliss 1946*, 27).

The geometric locus of the conjugate points P_1^* forms a curve that is the enveloping curve of the family of extremals E (solutions of the Euler-Lagrange equation). As soon as a curve of the family touches the enveloping curve belonging to the family, it is possible to use as well pieces of the envelope to solve the corresponding boundary value problem of the Euler-Lagrange equation. In other words, the extremity in these cases is lost. A vivid and concise discussion can be found in *Bliss 1946*, §§10–13. The necessary condition that the endpoint P_2 of an extremal E (here a geodesic) must come before the conjugate point P_1^* to the initial point of the extremal P_1 (with certain exceptions $P_2 = P_1^*$; see *Kneser 1898*, 50; *Osgood 1901a*, 166) is called the Jacobi condition.

Conjugate points limit (relative) extrema and thereby a fortiori also absolute extrema. Gaston Darboux had earlier emphasized, in his *Leçons sur la théorie générale des surfaces* (*Lectures on the general theory of surfaces*) (*1894*, 89), that the absolute extremum would always end before the relative extremum, or more precisely, that the points P_1^{**} that determine the end of the absolute extremum on the extremal E lie on E before those points P_1^* that determine the end of the relative extremum (therefore within the arc $P_1P_1^*$ of E). Analogously to the geometric locus of the conjugate point P_1^* to the initial point P_1 of the extremal, the enveloping curve of the extremal family, Zermelo introduces for the "by far shortest" paths (that is, for absolute extrema of the shortest paths) a curve, which he calls a *Doppelabstandskurve* (*double distance curve*); it presents the general situation for geodesics mentioned by Darboux in a more precise form.

Zur Theorie der kürzesten Linien

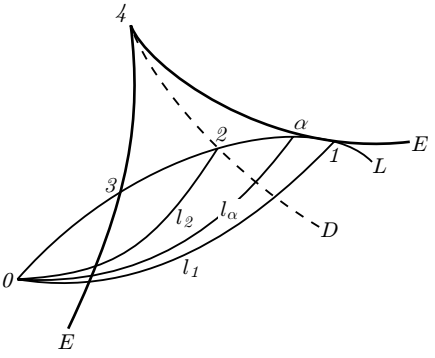
1902d

Das Problem der kürzesten Linien auf einer Fläche kann als das klassische Problem der Variationsrechnung bezeichnet werden, die zu seiner Lösung verwendeten Methoden sind vorbildlich und charakteristisch für die Methoden der Variationsrechnung überhaupt. Im Folgenden sei es mir daher gestattet, auf einige *Erweiterungen* dieses Problems hinzuweisen, die in ihrer anschaulichen Form zur Aufklärung allgemeinerer Variationsprobleme beitragen können.

1. Zunächst unterscheide ich zwischen *„kürzesten“* und *„allerkürzesten“* Linien, je nachdem sie nur in einem gewissen Flächenstücke ihrer Umgebung oder auf der ganzen Fläche die kürzeste Verbindung ihrer Endpunkte darstellen.

Jede von einem Punkte 0 ausgehende geodätische Linie L ist kürzeste bis zum nächsten „konjugierten Punkt“ 1, in dem sie die zu 0 gehörende Enveloppe E berührt. Aber schon zwischen 0 und 1 ist sie (nach *Kneser*) keine kürzeste mehr, und die allerkürzeste Verbindung beider Punkte ist eine von L verschiedene zweite geodätische Linie von der Länge $l_1 < L_1$. Dann werden auch die an l_1 sich stetig anschließenden geodätischen Linien l_α, deren Endpunkte α sehr nahe | vor 1 auf L liegen, immer noch kürzer sein als die auf L gemessenen geodätischen Abstände L_α dieser Punkte vom Anfangspunkt. Die geodätische Linie L hört also schon *vor* ihrem ersten konjugierten Punkte auf, eine „allerkürzeste“ zu sein, und es giebt eine Übergangsstelle 2, wo sie von einer gleichlangen geschnitten wird, sodaß $L_2 = l_2$ ist. Der geometrische Ort dieser Übergangspunkte 2 bildet die zu 0 gehörige „Doppelabstandskurve“ D, in der immer zwei gleich lange von 0 ausgehende kürzeste sich schneiden. Jede geodätische Linie ist also allerkürzeste nur bis zu einem gewissen Schnittpunkt mit der Doppelabstandskurve, der immer *vor* ihrem Berührungspunkt mit der Enveloppe liegen muß. (Für den Fall, daß sich in 0 keine kürzeste Linie selbst durchschneidet, können die geodätischen Linien l_α nicht kürzeste bleiben, wenn α an 0 heranrückt, sondern es giebt einen Grenzpunkt 3, der

185



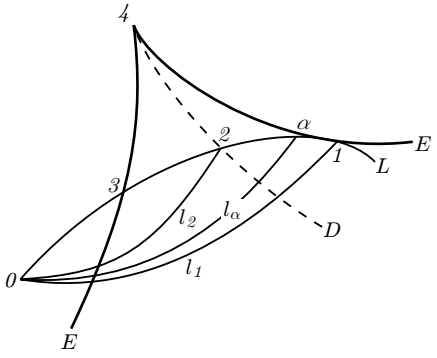
On the theory of shortest lines

1902d

The problem of the shortest lines on a surface may be considered the classic problem of the calculus of variations. The methods used for solving it are exemplary for and characteristic of the methods of the calculus of variations as a whole. Therefore, in what follows I may be permitted to suggest several *extensions* of this problem that, due to their distinct form, can help elucidate more general variational problems.

1. I first distinguish between "*shortest*" and "*by far shortest*" lines, depending on whether they represent the shortest connection between their endpoints only in a certain surface region of their neighborhood or on the entire surface.

Every geodetic line L starting from a point 0 is the shortest one up to the next "conjugate point" 1 at which it makes contact with the envelope E belonging to 0. However, already between 0 and 1 it is no longer the shortest one (according to *Kneser*), and by far the shortest connection between the two points is a second geodetic line different from L of length $l_1 < L_1$. Then the geodetic lines l_α which are continuously attached to l_1 and whose terminal points α lie very close before 1 on L will still be shorter than the geodetic distances L_α, measured along L, of these points from the starting point. Hence, the geodetic line L ceases to be a "by far shortest" line already *before* its first conjugate point, and there exists a transitional position 2, where it is intersected by an equally long line so that we have $L_2 = l_2$. The geometric location of this transitional point 2 forms the "double distance curve" D belonging to 0 in which two shortest lines of equal length starting from 0 always intersect. Hence, every geodetic line is a by far shortest line only up to a certain point where it intersects with the double distance curve and which must always lie *before* its point of contact with the envelope. (In the case where no shortest line intersects itself in 0 the geodetic lines l_α cannot remain shortest lines if α approaches 0. Instead, there then exists a boundary



auf l_3 die Rolle des konjugierten spielt und einen Schnittpunkt der Linie L mit der Enveloppe E darstellt.) Schließlich geht die Doppelabstandskurve D durch die Spitzen 4 der Enveloppe, in denen die entsprechenden Punkte 1, 2, 3 und die Linien L_4 und l_4 zusammenfallen.

2. Sucht man die kürzesten Linien innerhalb eines beliebig *begrenzten Flächenstückes*, so hat man ein Variationsproblem, in dem *Ungleichungen* als Nebenbedingungen auftreten. Ich beschränke mich hier auf ein einfach zusammenhängendes *ebenes* Flächenstück C, das also von einer geschlossenen sich selbst nicht schneidenden Kurve $\mathfrak{C}$ begrenzt ist, welche ihrerseits aus Stücken mit stetiger Tangente und endlicher Krümmung bestehen möge, im übrigen aber beliebig verlaufen kann. Im Innern von C liege ein Punkt 0, und gesucht werden die kürzesten ganz innerhalb C verlaufenden Verbindungslinien L_x dieses Punktes 0 mit allen Punkten x der Peripherie, also die verschiedenen Lagen, die ein in 0 befestigter gespannter Faden annimmt, wenn man ihn über die ganze Peripherie $\mathfrak{C}$ herumführt, die für den Faden als undurchdringlich vorausgesetzt wird.

Jede solche kürzeste Linie setzt sich, wie aus der Theorie der ersten Variation bekannt ist, abwechselnd aus geradlinigen Stücken und aus Teilen der Peripherie zusammen, die *nach innen konvex* sind, und an den Übergangs-
186 stellen muß Berührung stattfinden. Nun läßt sich | zeigen, daß das System dieser Linien *stetig* über die ganze Peripherie fortgesetzt werden kann, und daß zwei solcher Linien wohl im Anfang ein Stück gemein haben, aber, einmal getrennt, sich niemals wieder schneiden können. Wäre nämlich das letztere der Fall, so würden die beiden kürzesten Linien zwischen den Punkten 1 und 2 ein Flächenstück C' einschließen, das ganz im Innern von C läge und dessen Begrenzung mit Ausnahme der beiden Endpunkte *nirgends konvex* wäre. Das ist aber unmöglich, da unter den Parallelen zu der Geraden 12, die das Flächenstück durchsetzen, immer eine *letzte* existieren und diese dann notwendig an einer *konvexen* Stelle 3 die Peripherie berühren müßte. Die gefundenen kürzesten Linien L_x sind also die einzigen und damit auch die allerkürzesten innerhalb C und zerlegen *fächerförmig* das ganze Flächenstück C. Durch eine beliebig kleine Abänderung dieser Figur könnte man nun die zusammenfallenden Linienstücke durch getrennte ersetzen und mit Hilfe des entstehenden Strahlensystems das ganze Innere des Flächenstückes C ein-eindeutig und stetig auf das Innere eines Kreises abbilden. Doch bedarf dies noch eines strengen Beweises.

3. Anstatt einfacher Ungleichungen kann man auch Ungleichheiten zwischen Differentialausdrücken, „Differentialungleichungen“, als Nebenbedingungen einführen, man kann z. B. fordern, daß die *Steilheit* der Kurve eine gegebene Grenze nicht überschreitet: $\varkappa = \dfrac{dz}{ds} \leqq \varkappa_0$, wenn die Fläche gegeben ist durch eine Gleichung: $z = \varphi(x, y)$. Man sucht also für zwei Punkte der Fläche die kürzeste unter allen Verbindungslinien von „begrenzter Steilheit“, ein Problem, wie es praktisch angenähert verwirklicht ist beim Bau von Gebirgsstraßen. Die Lösungen dieses Problems setzen sich, wie leicht ersichtlich,

point 3 that plays the role of the conjugate point along l_3 and represents an intersection point of the line L with the envelope E.) The double distance curve D eventually runs through the cusps 4 of the envelope in which the corresponding points 1, 2, 3 and the lines L_4 and l_4 coincide.

2. If we seek the shortest lines in the interior of an arbitrarily *bounded surface region*, then we are confronted with a variational problem involving *inequalities* as auxiliary conditions. I shall restrict myself here to a simply connected *planar* surface region C, and hence to one that is bounded by a closed curve $\mathfrak{C}$ not intersecting itself which, in turn, may consist of regions with continuous tangent and finite curvature but whose course is otherwise arbitrary. Let 0 be a point in the interior of C. Now find those shortest lines L_x that lie entirely within C and connect this point 0 with all points x of the periphery, and hence the various positions taken by a taut thread fixed at 0, if we pass it round the entire periphery $\mathfrak{C}$, which we take to be impenetrable to the thread.

As we know from the theory of the first variation, each such shortest line consists of alternately rectilinear portions and of parts of the periphery that are *inwardly convex*, and contact must be made at the transitional positions. Now, it is possible to show that the system of these lines is capable of a continuous continuation across the entire periphery and that once separated any two such lines, while certainly capable of sharing an initial segment, can never intersect one another again. For if the latter were the case, then the two shortest lines between the points 1 and 2 would enclose a surface region C' which would lie entirely in the interior of C and whose boundary except for the two terminal point would be *nowhere convex*. This, however, is impossible since among the lines parallel to the straight line 12 that lace the surface region there would always have to be a *last one* which then, by necessity, would have to make contact with the periphery at a *convex* point 3. The shortest lines L_x found are therefore the only ones in existence, and hence also the by far shortest lines within C, and they divide the entire surface region C *like a fan*. By an arbitrarily small modification of this figure, we could now replace the coinciding line segments by separate ones and, by means of the emerging system of rays, map one-to-one and continuously the entire interior of the surface region C onto the interior of a circle. This, however, still stands in need of rigorous proof.

3. Instead of simple inequalities, we can also introduce inequalities between differential expressions, "differential inequalities", as auxiliary conditions. We can, e.g., require that the *steepness* of the curve not exceed a given limit: $\varkappa = \dfrac{dz}{ds} \leqq \varkappa_0$, if the surface is given in terms of an equation: $z = \varphi(x, y)$. Hence, what is sought is, given two points of the surface, the shortest among all connecting lines of "limited steepness", a problem that is approximately realized in the construction of mountain roads. As is readily evident, the solutions to this problem are composed of shortest *geodetic*

zusammen aus kürzesten *geodätischen* Linien, soweit sie nirgends zu steil sind, und aus Kurven von *konstanter Steilheit* $\varkappa = \varkappa_0$, und an den Übergangsstellen müssen auch die geodätischen Linien dieselbe maximale Steilheit besitzen. Jedes Stück einer solchen Linie konstanter Steilheit, die ich zur Abkürzung einfach als „Kletterkurve“ bezeichnen will, ist nun auch wirklich immer die kürzeste Verbindung seiner Endpunkte unter allen denen, die nirgends steiler sind, und zwei Kletterkurven von derselben Steilheit zwischen denselben Punkten sind auch immer gleich lang. Da es nun in jedem Flächenpunkte im allgemeinen zwei reelle oder imaginäre Richtungen giebt, für welche die Steilheit einen gegebenen Wert $\varkappa$ annimmt, so giebt es zwei einfach-unendliche Scharen von Kletterkurven $\varkappa_0$, die alle Teile der Fläche, in denen diese Steilheit $\varkappa_0$ überhaupt möglich ist, netzförmig überdecken und an den Grenz-
187 | kurven dieser Gebiete in Spitzen endigen. In einem solchen Gebiete kann man nun zwei Punkte, die nicht auf derselben Kletterkurve liegen, durch abwechselnde Verwendung der beiden Scharen gleichwohl sehr häufig durch eine Linie konstanter Steilheit $\varkappa_0$ verbinden, also durch eine Zickzacklinie oder Serpentine, wie wir sie auf den meisten Gebirgspässen beobachten können. Nun läßt sich zeigen: jede Kurve von überall *größerer* Steilheit $\varkappa > \varkappa_0$ kann immer durch eine zickzackförmige Kletterkurve $\varkappa_0$ ersetzt werden, die in beliebiger Nähe zwischen denselben Endpunkten verläuft und allen Bedingungen des Problems genügt. Solche Kurven $\varkappa > \varkappa_0$ giebt es aber von einem Anfangspunkte O aus nach allen Punkten eines Gebietes, das von den beiden von O ausgehenden einfachen Kletterkurven und außerdem von der Grenzkurve des betreffenden Steilheitsgebietes oder des Kletterkurven-Netzes $\varkappa_0$ begrenzt wird.

Wollte man dagegen, wie beim Eisenbahnbau, plötzliche Richtungsänderungen ausschließen und auch für die *Krümmung* des Weges eine obere Grenze k_0 festsetzen, so wäre das Problem in vielen Fällen überhaupt unlösbar, weil es solche Kurven zwischen den betrachteten Punkten auf der Fläche überhaupt nicht gäbe. In solchen Fällen müßte man notwendig die Fläche verlassen und in den Raum eindringen, also, technisch gesprochen, Tunnels und Viadukte anlegen. Dabei ergeben sich denn auch unter anderem die schraubenförmigen „Kehrtunnels“ verschiedener Gebirgsbahnen als mathematisch korrekte Lösungen des Variationsproblems.

lines, provided that they are nowhere too steep, and of curves of *constant steepness* $\varkappa = \varkappa_0$, and at the transitional positions the geodetic lines, too, must have the same maximal steepness. Now, every segment of such a line of constant steepness, or "climbing curve" for short, is really always the shortest connection of its endpoints among all those that are nowhere steeper, and two climbing curves of identical steepness between the identical points are also always of the same length. Since now in every surface point there are, in general, two real or imaginary directions for which the steepness takes a given value $\varkappa$, there are two simply-infinite families of climbing curves $\varkappa_0$ that cover all parts of the surface in which this steepness $\varkappa_0$ is possible at all like a net and that terminate in cusps at the boundary curves of these regions. Now, in such a region, it is often possible to connect two points that do not lie on the same climbing curve by a line of constant steepness $\varkappa_0$ by alternately using the two families, and hence by a zig-zag line or serpentine, as we can find them at most mountain passes. We can now show: every curve of everywhere *greater* steepness $\varkappa > \varkappa_0$ can always be replaced by a zig-zag-like climbing curve $\varkappa_0$ that runs at arbitrary proximity between the same endpoints and meets all conditions of the problem. Such curves $\varkappa > \varkappa_0$ exist, however, running from a starting point O to all points of a region that is bounded by the two simple climbing curves starting from O and, in addition, also by the boundary curve of the respective steepness region or of the climbing curve net $\varkappa_0$.

If, in contrast, sudden changes in direction were to be excluded, as in the case of railway construction, and an upper limit k_0 were to be fixed also for the *curvature* of the path, then, in many cases, the problem would be simply unsolvable, for such curves would simply not exist between the considered points on the surface. In such cases, we would, by necessity, have to leave the surface and move into space, and hence, technically speaking, build tunnels and overpasses. Here, we find as mathematically correct solutions to the variational problem also the coil-like "helical tunnels" of various mountain railways.