J. Austral. Math. Soc. (*Series A*) **64** (1998), 37–50

A GLOBAL ALGORITHM FOR GEODESICS

LYLE NOAKES

(Received 30 September 1997; revised 7 November 1997)

Communicated by R. Bartnik

Abstract

The problem of finding a geodesic joining given points x_0, x_1 in a connected complete Riemannian manifold requires much more effort than determining a geodesic from initial data. Boundary value problems of this type are sometimes solved using shooting methods, which work best when good initial guesses are available, especially when x_0, x_1 are nearby. Galerkin methods have their drawbacks too. The situation is much more difficult with general variational problems, which is why we focus on the Riemannian case.

Our global algorithm is very simple to implement, and works well in practice, with no need for an initial guess. The proof of convergence is elementary and very carefully stated, with a view to possible generalizations later on. We have in mind the much larger class of interesting problems arising in optimal control especially from mechanical engineering.

1991 *Mathematics subject classification (Amer. Math. Soc.)*: primary 34B15, 49M05; secondary 53C22.

1. Introduction

Let N be a C^∞ path-connected Riemannian n-manifold where $n \geq 1$ is finite. When N is complete with respect to the Riemannian distance function d the classical Hopf-Rinow theorem says that any $x_0, x_1 \in N$ are joined by a minimal geodesic, namely a curve $\gamma : [0, 1] \to N$ of minimum length with respect to the Riemannian structure. When the geometry of N is very well understood all geodesics can be written down in closed form [1, 6], but in general finding γ is not easy.

There always exists a coordinate chart of N containing the image of γ, but finding the chart may not be easy either, unless γ is given or x_0, x_1 are nearby. Putting that difficulty to one side, γ solves a second order non-linear system of n ordinary differential equations defined in the chart coordinates, as well as the boundary conditions $\gamma(i) = x_i$ for $i = 0, 1$. So the search for γ can be considerably narrowed by solving a

2-point boundary value problem, at least if a suitable coordinate chart has been found.

Solving a 2-point boundary-value problem is much harder than solving an initial-value problem. In particular, some kind of completeness assumption, such as we have made in the Riemannian case, is needed to ensure that a solution to the initial-value problem exists for all time. Typically the boundary-value problem does not have a unique solution, and in the general case a solution need not exist. In the single shooting method [3, Chapter 2] for 2-point boundary-value problems the unknown initial data is estimated, or just guessed when no basis exists for making an estimate. Then the corresponding initial-value problem is solved to obtain an estimate $\hat{\gamma}$ of the solution to the boundary-value problem. The error $\hat{\gamma}(1) - x_1$ in the terminal value is used to update the initial guess. When the initial guess is good the estimates obtained by iterating this procedure converge to a solution of the boundary-value problem. The success rate does not seem to be high in other cases, and in general there is no guarantee of convergence.

When x_0, x_1 are nearby they determine a useful estimate of the initial velocity $\dot{\gamma}(0)$, which is the extra initial data that we need for single-shooting. Usually in such cases, which we call the *local* version of our problem, single shooting works well and is possibly the method of choice. When x_0, x_1 are distant we have the *global* problem and the performance of single-shooting is critically dependent on the quality of the initial guess.

Error accumulation in solutions of initial-value problems can be especially troublesome for non-linear systems. To cope with this and other difficulties, single shooting is sometimes replaced by *multiple shooting* [3]. Then $[0, 1]$ is divided into small subintervals whose initial data is simultaneously updated at each step. This ameliorates chaotic effects but performance is still heavily dependent on the quality of the initial guess. As an added computational burden, the number of variables is substantially increased. As for single shooting there is usually no guarantee of convergence.

The global algorithm of the present paper resembles multiple shooting in that $[0, 1]$ is subdivided and geodesics are found separately over each subinterval. So we do not expect to be troubled by the non-linear dynamics either. The most important differences between the global algorithm and multiple shooting, as described in [3], are

(1) Each step of the global algorithm updates only n real variables at a time.

(2) The curves of the global algorithm satisfy both boundary conditions at every step of the iteration.

(3) The global algorithm always converges, without the need for an initial guess. Usually there is no need to search for convergent subsequences. The entire sequence of approximations is proved to converge under fairly general conditions.

The present paper exploits the success of single shooting by treating the local

problem as essentially solved. This opens up the possibility of solving the global problem by building approximations from local solutions. This idea is certainly not new. For example in [6, III.Section 16] the space of all piecewise-C^1 curves joining x_0, x_1 is approximated by a C^∞ finite-dimensional manifold B of piecewise-geodesics. Restricting the energy integral E to a suitable compact subset of B suffices to prove the Hopf-Rinow Theorem. So it appears we might be on the right track.

Indeed the method of gradient descent applied to the C^∞ function $E' \equiv E|B$ on B can be used to solve the global problem. Alternatively, gradient descent can be applied directly to suitable infinite dimensional manifolds of curves [11]. The gradient of E' is a vector of velocity increments at the junctions of a piecewise geodesic and is readily calculated in practice. However a practical difficulty with gradient descent is that each iteration requires a choice of step-size. The most satisfactory way to make the choice is to base it on a local quadratic approximation to E' namely by reference to the Hessian H of E'. Usually H does not need to be updated at every step, but the need to calculate it at all substantially increases the computational effort required. The dimension b of B is nj where j is the number of junctions in the piecewise-geodesics. When step-sizes are determined by human intervention, gradient descent takes place in $\mathbb{R}^b$. Once the process is fully automated H represents a further $b(b+1)/2$ scalars.

The Gauss-Seidel algorithm is an iterative scheme for the solution of large systems of affine equations. Each iteration adjusts a single variable, and for large systems Gauss-Seidel is much more efficient than Gaussian elimination. We are faced with not dissimilar difficulties in the application of gradient descent to the global problem, especially when b is large. So it seems natural to imitate Gauss-Seidel. (An alternative way of motivating the global algorithm is by comparison with the non-linear corner-cutting techniques of [7–9].)

Consider a piecewise-geodesic curve $\omega : [0, 1] \to N$ from x_0 to x_1, parameterized proportionally to arc-length, and whose j geodesic segments occur within convex subsets of N. Then ω is uniquely defined by the j-tuple $(y_1, y_2, \ldots, y_j) \in N^j$ of junctions of geodesic segments. Instead of applying gradient descent to the nj-dimensional j-tuple we adjust each y_i separately as follows. Set $y_0 = x_0, y_q = x_1$ where $q = j + 1$ and suppose that for each $1 \leq i < q$ all three of y_{i-1}, y_i, y_{i+1} lie inside some convex subset of N. Then moving y_i onto the minimal geodesic joining y_{i-1}, y_{i+1} achieves the largest possible decrease in length while keeping other variables fixed. There is some uncertainty about where on the minimal geodesic y_i should go but in order to focus the discussion we settle on the midpoint. The global algorithm consists of iterating this procedure so that all y_i are moved infinitely often where $0 < i < q$. More precisely, in the present paper let i run from 1 to $q - 1$ and then start over again. This generates a sequence $\Omega = \{\omega_a : [0, 1] \to N : a \geq 1\}$ of piecewise-geodesics whose lengths are decreasing.

A little attention to detail shows that Ω has a uniformly convergent subsequence. It

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

is plausible, true, but not quite obvious that the limit γ is a geodesic. (However γ need not be a minimal geodesic.) What complicates the proof a little is that the $(q+1)$-tuple determining γ might contain redundancies. So we obtain a useful algorithm which is not very demanding of computational resources.

Even greater efficiencies are possible when Ω is known to be convergent, because then we do not have to look out for convergent subsequences. We prove that Ω converges when N has everywhere non-positive sectional curvature, and in many other situations as well. We do not know whether Ω is *always* convergent. The only case where we might have to go to subsequences is where there are distinct geodesics $\gamma_i : [0, 1] \to N$ of the same length, which are homotopic through curves from x_0 to x_1.

EXAMPLE 1.1. If $N = S^n$ with the usual Riemannian metric then Ω converges unless $x_0 = -x_1$. If $x_0 = -x_1$ we might have to go to subsequences. Finding geodesics on S^n is no trouble at all because the geometry is so well understood. The same goes for the next two examples.

EXAMPLE 1.2. If $N = \mathbb{R}P^n$ with the usual Riemannian metric then Ω converges unless $d(x_0, x_1) = \pi$.

EXAMPLE 1.3. If N is the n-dimensional flat torus $S^1 \times S^1 \times \cdots \times S^1$ then Ω converges for any choice of x_0, x_1.

Before going into the details of the global algorithm we mention an alternative and very attractive method of constructing solutions to non-linear variational problems, namely the use of pseudomonotone operators in non-linear functional analysis. Except for the difficulty already mentioned (which is not a problem for the global algorithm) of finding a suitable coordinate chart, the problem of joining x_0, x_1 by a geodesic can also be approached using a very general result of Brézis [16, Theorem 27.A]. Apart from its wide range of possible applications, for us the most interesting aspect of this theorem is that it gives an effective construction using a sequence of Galerkin approximations. The approximations are found by solving a non-linear system of equations in $\mathbb{R}^k$ for increasing values of k, appealing to the Brouwer fixed point theorem each time. Turning the Brouwer theorem into a constructive procedure is not without its practical difficulties, and of course this has to be carried out time after time as k increases.

A more serious difficulty with the Galerkin approximations is that k is unbounded, namely, there is an explosion in the number of variables that need to be considered. So Brézis' very important theorem seems somewhat deficient as a practical method of solving variational problems and in particular for finding geodesics. The global algorithm does not suffer from the same limitations because the space B of approximating

curves actually contains the geodesic and its dimension is $n(q-1)$ where q depends on the geometry of N.

Galerkin approximations have nonetheless been used to achieve impressive successes in solving practical problems in optimal control, for example in the work of Teo and his co-workers [14]. For instance there is no doubt at all that the MISER software package can find geodesics. However it must also be admitted that the computational effort required for these successes is sometimes very large, as would be for an implementation of the proof of Brézis' theorem. It was to ease this computational burden that a version of the global algorithm for optimal control was proposed in stimulating conversations between K. L. Teo, the present author, and their research student Y. C. Liu. These conversations were motivated in part by the work of Zuo [17] on an algorithm for discrete-time optimal control problems, but our efforts were soon abandoned due to difficulties of proving (or even verifying) convergence. The algorithm in the present paper is a continuous-time analogue of Zuo's. The present author will revisit optimal control in future papers. However there are some challenging problems calling for the computation of geodesics. One problem which seems more accessible now, in light of the present paper, is the following

EXAMPLE 1.4. In the statistical problem of computing Rao distances between multivariate Gaussian distributions with different means [5, 12, 13] the Riemannian distances are very difficult to compute in closed form. There is a single exception, reducing to planar hyperbolic geometry.

For future work we have in mind the much larger class of interesting problems arising in optimal control especially from mechanical engineering [2, 10].

2. Midpoint maps

A subset W of N is said to be *convex* when

(1) for any $x_0, x_1 \in W$, there is a minimal geodesic $\gamma : [0, 1] \to W$ of N from x_0 to x_1;

(2) γ is the only geodesic from x_0 to x_1 defined on $[0, 1]$ whose image is entirely contained in W;

(3) γ depends differentiably on x_0, x_1.

By [15] N has an open cover by convex sets.

Let $L(\omega)$ be the Riemannian length of a piecewise-C^1 curve $\omega : [a, b] \to N$. Let d be the Riemannian distance function namely the metric d on N given by

$$d(x_0, x_1) = \inf\{L(\omega)\}$$

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

where $\omega : [0, 1] \to N$ varies over piecewise-C^1 curves from x_0 to x_1. Let the metric space (N, d) be complete. Then the closure of any open ball $B(x_0, r)$ is compact.

Given a piecewise-C^1 curve $\omega : [0, 1] \to N$, let $6\delta > 0$ be a Lebesgue number of an open cover $\{W_\alpha : \alpha \in A\}$ of the closure of $B(\omega(0), L(\omega))$ by convex subsets of N. Let X be the union $\bigcup_{\alpha \in A} W_\alpha$.

Let $D = \{(x_0, x_1) \in X \times X : d(x_0, x_1) \leq 2\delta\}$ and define $M : D \to N$ by $M(x_0, x_1) = \gamma(1/2)$ where $\gamma : [0, 1] \to N$ is the minimal geodesic from x_0 to x_1. Because of the following simple result M maps into X.

LEMMA 2.1. *For $(x_0, x_1) \in D$, $M(x_0, x_1) \in X$.*

PROOF. Because $d(x_0, x_1) < 3\delta$, $x_0, x_1 \in W_\alpha$ for some $\alpha \in A$. Then $M(x_0, x_1) \in W_\alpha$ because W_α is convex. This proves the lemma.

Now M is C^∞ and

$$d(x_0, M(x_0, x_1)) = d(x_0, x_1)/2 = d(M(x_0, x_1), x_1) \tag{1}$$

for all $(x_0, x_1) \in D$.

Choose $0 = t_0 < t_1 < \cdots < t_q = 1$ so that $d(\omega(t_{i-1}), \omega(t_i)) \leq \delta$ for all $i = 1, 2, \ldots, q$. Because each $L(\omega|[0, t_i]) \leq L(\omega)$, $\omega(t_i) \in X$. In other words $(\omega(t_0), \omega(t_1), \ldots, \omega(t_q))$ is an element of the set Y of all $(q + 1)$-tuples $y = (y_0, y_1, \ldots,$
$y_q) \in X^{q+1}$ satisfying $d(y_{i-1}, y_i) \leq \delta$ for all $i = 1, 2, \ldots, q$.

For $1 < p < q$ define $G_p : Y \to X^{q+1}$ by

$$G_p(y) = (y_0, y_1, \ldots, y_{p-1}, z_p, y_{p+1}, y_{p+2}, \ldots, y_q)$$

where $z_p = M(y_{p-1}, y_{p+1})$.

LEMMA 2.2. *$G_p(y) \in Y$.*

PROOF.

$$\begin{aligned} d(y_{p-1}, z_p) &= d(z_p, y_{p+1}) = d(y_{p-1}, y_{p+1})/2 \\ &\leq (d(y_{p-1}, y_p) + d(y_p, y_{p+1}))/2 \leq (2\delta)/2. \end{aligned}$$

The lemma is proved.

So $G_p : Y \to Y$ where $0 < p < q$. Since M is C^∞ so are the G_p. Define a C^∞ function $F : Y \to Y$ as the composite $G_{p-1} \circ G_{p-2} \circ \cdots \circ G_1$. Then $F(y) \in Y$ is the $(q + 1)$-tuple z defined inductively by

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

(1) $z_0 = y_0$,
(2) $z_i = M(z_{i-1}, y_{i+1})$ for $1 \le i < q$,
(3) $z_q = y_q$.

Note that $F(y)$ does not depend on y_1.

Just as the piecewise-C^1 curve ω has length so does $y \in Y$: define $\lambda(y) = \Sigma_{i=1,2,\ldots,q} d(y_{i-1}, y_i)$. Then $d(y_0, y_q) \le \lambda(y) \le L(\omega)$ and $\lambda : Y \to \mathbb{R}$ is continuous.

LEMMA 2.3. *For $0 < p < q$ and all $y \in Y$, $\lambda(G_p(y)) \le \lambda(y)$.*

PROOF. $\lambda(y) - \lambda(G_p(y))$ is

$$\begin{aligned} &d(y_{p-1}, y_p) + d(y_p, y_{p+1}) - d(y_{p-1}, z_p) - d(z_p, y_{p+1}) \\ &\quad = d(y_{p-1}, y_p) + d(y_p, y_{p+1}) - d(y_{p-1}, y_{p+1}) \ge 0. \end{aligned}$$

This proves the lemma.

Thus $q - 1$ applications give

LEMMA 2.4. *For all $y \in Y$, $\lambda(F(y)) \le \lambda(y)$.*

Let d^{q+1} be the uniform metric on Y induced by d, namely:

$$d^{q+1}(y, z) = \max_{i=0,1,\ldots,q} d(y_i, z_i).$$

LEMMA 2.5. $d^{q+1}(y, F(y)) \le 2\delta$.

PROOF. Write $z = F(y)$. Then $d(y_0, z_0) = d(y_q, z_q) = 0$. For $0 < i < q$

$$d(z_i, y_{i+1}) = d(z_{i-1}, z_i) \le \delta$$

because $z \in Y$. But $y \in Y$ also and so $d(y_i, y_{i+1}) \le \delta$. Therefore $d(y_i, z_i) \le 2\delta$ and this proves the lemma.

3. Multiplicities and curves

So as to simultaneously study $(q+1)$-tuples for different values of q, the notation of Section 2 is supplemented when necessary with superscripts (q) referring to $(q+1)$-tuples. So $Y^{(q)}$ is the space Y defined in Section 2 and $Y^{(p)}$ is the same but with $(p+1)$-tuples instead of $(q+1)$-tuples. The same symbol is used for F in the context of $(q+1)$-tuples regardless of the value of q.

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

DEFINITION 3.1. y has *multiplicity* $\geq k - j + 1$ in position $0 < j < q$ when $y_j = y_{j+1} = \cdots = y_k$. The *multiplicity* of y in position $0 < j < q$ is the largest $1 \leq m \leq q$ such that y has multiplicity $\geq m$ in position j.

The *reduction* $\rho(y)$ of $y \in Y^{(q)}$ is defined by discarding consecutive repetitions in positions $0 < j < q$. Note

(1) for some $0 < r \leq q$, $\rho(y) \in Y^{(r)}$ and has multiplicity 1 in every position $0 < j < r$;

(2) $\rho(y) = y$ if and only if y has multiplicity 1 in every position $0 < j < q$. In such a case y is said to be *irreducible*.

The *expansion* $\epsilon(w, m)$ of $w \in Y^{(r)}$ by an $(r+1)$-tuple $m = (m_0, m_1, \ldots, m_r)$ of positive integers is obtained by replacing each w_i with m_i copies of itself. Any $y \in Y$ can be written in the form $\epsilon(w, m)$ for some m, where w is the irreducible $\rho(y)$. The *curve* of a $(q+1)$-tuple $y \in Y$ is a piecewise-geodesic joining the y_i in order, and parameterised proportionally to arc-length. More precisely

DEFINITION 3.2. For $y \in Y$ and $0 < i \leq q$ let $\gamma_i : [0, 1] \to N$ be the minimal geodesic from y_{i-1} to y_i. The *curve* of $y \in Y$ is $\omega_y : [0, 1] \to N$ where

$$\omega_y(t) = \gamma_i \left(\left(t\lambda(y) - \sum_{0<j<i} d(y_{j-1}, y_j) \right) / d(y_{i-1}, y_i) \right)$$

and

$$t\lambda(y) \in \left[\sum_{0<j<i} d(y_{j-1}, y_j), \sum_{0<j\leq i} d(y_{j-1}, y_j) \right].$$

Note the following simple consequences of the definitions of ω_y, λ and L.

$$L(\omega_y) = \lambda(y), \tag{2}$$

$$\omega_{\rho(y)} = \omega_y. \tag{3}$$

From now until the end of this section ω_y will be a geodesic $\gamma : [0, 1] \to N$. This exceptional situation occurs in the proof of Lemma 4.2, as the result of a limiting process. In such a case we have

$$\omega_{F(y)} = \omega_y \tag{4}$$

We can write $F(y) = z$ where $z_i = \gamma(u_i)$ and $u_0 = 0$, $u_i = (u_{i-1} + t_{i+1})/2$ for $0 < i < q$, $u_q = 1$.

LEMMA 3.1. *When ω_y is a geodesic, with u_i, t_i as above,*

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

(1) $u_i \le t_{i+1}$ *for* $0 \le i < q$;
(2) $u_i \le u_{i+1}$ *for* $0 \le i < q$;
(3) *if* $t_{j-1} < t_j = t_{j+1} = \cdots = t_k$ *then* $u_i < t_{i+1}$ *for* $j-1 \le i < k$.

PROOF. We first prove that $u_i \le t_{i+1}$ by induction on $0 \le i < q$. When $i = 0$ we have $u_0 = 0$ and $t_1 \ge t_0 = 0$. For $i > 0$ suppose inductively that $u_{i-1} \le t_i$. Then $u_i \le (t_i + t_{i+1})/2 \le t_{i+1}$ and so the assertion is proved.

The second assertion holds trivially when $i = q-1$. For $0 \le i < q-1$ we have

$$u_{i+1} \ge (u_i + t_{i+2})/2 \ge (u_i + t_{i+1})/2 \ge u_i$$

according to the first assertion.

The last part of the lemma is proved by induction. When $j = 1$, $u_0 = 0 < t_1$ by assumption. When $i = j-1 > 0$, $u_{j-1} = (u_{j-2} + t_j)/2 \le (t_{j-1} + t_j)/2$ according to the first assertion. Since $t_{j-1} < t_j$ we obtain $u_{j-1} < t_j$.

Now for $j-1 < i < k$ suppose inductively that $u_{i-1} < t_i$. Then $u_i < (t_i + t_{i+1})/2 = t_{i+1}$ which completes the proof of the lemma.

The following result is not used in the present paper. It is included for completeness.

LEMMA 3.2. *Let* ω_y *be a geodesic* γ. *Then the sequence* $\{F^a(y) : a \ge 1\} \subset Y$ *converges to the uniformly distributed* $(q+1)$*-tuple*

$$(y_0, \gamma(1/q), \gamma(2/q), \ldots, \gamma(i/q), \ldots, y_q).$$

PROOF. If $q = 2$ the result is clear since $F(y)$ is independent of y_1 and the limit is achieved immediately as $F(y)$. For $q > 2$ write

$$t = (t_2, t_3, \ldots, t_{q-1})^T, \qquad u = (u_2, u_3, \ldots, u_{q-1})^T \in \mathbb{R}^{q-2}.$$

If $F(y) = z$ then $t = Au + b$ where A is the $(q-2) \times (q-2)$ matrix whose rows are

$$\begin{matrix} 1/2^2 & 1/2 & 0 \cdots 0 \\ 1/2^3 & 1/2^2 & 1/2 & 0 \cdots 0 \\ & \cdots & & \\ 1/2^{q-1} & 1/2^{q-2} & 1/2^{q-3} & \cdots 1/2^2 \end{matrix}$$

and $b = (0, 0, \ldots, 0, 1/2)^T \in \mathbb{R}^{q-2}$. It follows that $F^a(y)$ is

$$(y_0, \gamma(u_2^{(a-1)}/2), \gamma(u_2^{(a)}), \gamma(u_3^{(a)}, \ldots, \gamma(u_{q-1}^{(a)}), y_q)$$

where $u^{(0)} = t$ and, for $a > 0$,

$$u^{(a)} = b + Ab + A^2 b + \cdots + A^{a-1} b + A^a t.$$

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

This converges since $\|A\| < 1$, and the limit $u^{(\infty)}$ satisfies $u^{(\infty)} = Au^{(\infty)} + b$. It is readily verified that $u^{(\infty)} = (2, 3, \ldots, q-1)^T/q$ is the unique solution, and this proves the lemma.

4. Extremes

Let $y \in Y$ and define $s^{(a)} = F^a(y)$ for $a \geq 1$. By Lemma 2.4 the sequence $\{\lambda(s^{(a)}) : a \geq 1\}$ converges to its infimum $\lambda^{(\infty)} \in [d(y_0, y_q), q\delta]$.

Because (N, d) is complete the closure B of $B(y_0, q\delta)$ is compact. Because Y is closed in the cartesian power B^{q+1}, Y is compact. Let $\{s^{(a_j)} : j \geq 1\}$ be any convergent subsequence of $\{s^{(a)} : a \geq 1\} \subset Y$. Write $\lim_{j\to\infty} s^{(a_j)} = s^{(\infty)} \in Y$, where $s^{(\infty)} = (s_0^{(\infty)}, s_1^{(\infty)}, \ldots, s_q^{(\infty)})$. Because λ is continuous $\lambda(s^{(\infty)}) = \lambda^{(\infty)}$.

Given $(x_0, x_1) \in D$ say that $w \in N$ is *between* x_0, x_1 when w lies in the image of the minimal geodesic $\gamma : [0, 1] \to N$ from x_0 to x_1. In such a case $(x_0, w), (w, x_1) \in D$. Note also the following.

(1) x_0 and x_1 are between x_0, x_1;
(2) w is between x_0, x_1 if and only if w is between x_1, x_0.

Call $w \in Y^{(p)}$ *extreme* when w_i is between w_{i-1}, w_{i+1} for all $1 < i < p$.

LEMMA 4.1. *An irreducible $w \in Y^{(p)}$ is extreme if and only if ω_w is a geodesic.*

PROOF. Suppose that w is extreme and for $0 < i < q$ let $\tilde{\gamma}_i : [0, 1] \to N$ denote the minimal geodesic from w_{i-1} to w_{i+1}. By contrast the minimal geodesics in Definition 3.2 from w_{i-1} to w_i are $\gamma_i : [0, 1] \to N$.

Because w_i is between w_{i-1}, w_{i+1}, $w_i = \tilde{\gamma}_i(s)$ where $s \in (0, 1)$ since w is irreducible. Then s is the ratio

$$d(w_i, w_{i+1})/(d(w_{i-1}, w_i) + d(w_i, w_{i+1})).$$

The restrictions $\tilde{\gamma}_i|[0, s]$ and $\tilde{\gamma}_i|[s, 1]$ are minimal geodesics, from w_{i-1} to w_i and from w_i to w_{i+1} respectively. So in Definition 3.2

$$\gamma_i(t) = \tilde{\gamma}_i(t/s) \qquad \text{and} \qquad \gamma_{i+1}(t) = \tilde{\gamma}_i((t-s)/(1-s)).$$

Consequently $s\dot{\gamma}_i(1) = (1-s)\dot{\gamma}_{i+1}(0)$, and substituting for s:

$$\dot{\gamma}_i(1)/d(w_{i-1}, w_i) = \dot{\gamma}_{i+1}(0)/d(w_i, w_{i+1}).$$

DEFINITION 4.1. Let a continuous curve γ be defined over a closed interval $[a, b]$ and suppose that $c \in (a, b)$. Then the restrictions of γ to the subintervals $[a, c]$ and $[c, b]$ are called *track-summands* of γ, and γ is the *track-sum* of its track-summands. More generally γ may be repeatedly decomposed into a track-sum of finitely many summands.

So the left and right derivatives of ω_w agree at the junctions of the track-sum, namely when $tL(w) = \Sigma_{0<j\le i} d(w_{j-1}, w_j)$ for $0 < i < p$. So ω_w is C^1, as well as a track-sum of geodesics. This proves that ω_w is a geodesic when w is extreme.

Suppose now that ω_w is a geodesic. For $0 < i < p$ the restriction $\tilde{\omega}_i$ of ω_w to the interval

$$\left[\sum_{0<j<i} d(w_{j-1}, w_j)/L(w), \sum_{0<j\le i+1} d(w_{j-1}, w_j)/L(w)\right]$$

is a geodesic from w_{i-1} to w_{i+1}. The diameter of the image of $\tilde{\omega}$ is at most 2δ and 3δ is a Lebesgue number of an open cover of N by convex sets. So the geodesic $\tilde{\omega}_i$ maps into some convex subset of N. So $\tilde{\omega}_i$ is a *minimal* geodesic from w_{i-1} to w_{i+1} and, after reparameterisation, w_i is seen to be between w_{i-1}, w_{i+1}. This completes the proof.

The first two parts of the following result come from Lemma 2.4 and the definition of F (they are included here only for convenience). The proof of the third assertion is complicated by the need to allow for the possibility that y might not be irreducible.

LEMMA 4.2. (1) $\lambda(F(y)) \le \lambda(y)$;
(2) *if* ω_y *is a geodesic then* $\lambda(F(y)) = \lambda(y)$;
(3) *if* $\lambda(F(y)) = \lambda(y)$ *then* ω_y *is a geodesic.*

PROOF. To prove the third assertion write $y = \epsilon(w, m)$ where $w = (w_0, w_1, \dots, w_r) \in Y^{(r)}$ is irreducible and m is an $(r+1)$-tuple of positive integers. Suppose ω_y is not a geodesic. Then w is not extreme by (3) of Section 3 and Lemma 4.1. Let $1 < p < r$ be the largest integer such that $\tilde{w} \equiv (w_0, w_1, \dots, w_p) \in Y^{(p)}$ is extreme. Then $\tilde{w}' \equiv (w_0, w_1, \dots, w_p, w_{p+1})$ is not extreme. By Lemma 4.1, $\omega_{\tilde{w}}$ is a geodesic and $\omega_{\tilde{w}'}$ is not. Note that $\omega_{\tilde{w}'}$ is a track-sum of $\omega_{\tilde{w}}$ with a minimal geodesic γ from w_p to w_{p+1}. For $0 \le i \le k = \Sigma_{j=1,\dots,p} m_j$ write $y_i = \omega_{\tilde{w}}(t_i)$ where

$$0 = t_0 \le t_1 \le t_2 \le \cdots \le t_{j-1} < t_j = t_{j+1} = \cdots = t_k = 1.$$

Write $\tilde{y} = (\omega_{\tilde{w}}(t_0), \omega_{\tilde{w}}(t_1), \dots, \omega_{\tilde{w}}(t_k))$. Then $\rho(\tilde{y}) = \tilde{w}$, and $\omega_{\tilde{y}} = \omega_{\tilde{w}}$ by (3). So the track-sum of $\omega_{\tilde{y}}$ with γ is not a geodesic. Then by (4), (3) the track-sum of $\omega_{\rho(F(\tilde{y}))}$ with γ is not a geodesic.

By Lemma 3.1, $F(\tilde{y}) = (y_0, z_1, \dots, z_{j-1}, z_j, \dots, z_{k-1}, w_p)$ where $z_i = \omega_{\tilde{w}}(u_i)$ for $0 < i < k$, and

(1) $u_i \le t_{i+1}$ for $0 \le i < k$;
(2) $u_i \le u_{i+1}$ for $0 \le i < q$;
(3) $u_i < t_{i+1}$ for $j - 1 \le i < k$.

In particular $u_{k-1} < t_k = 1$ and $z_{k-1} \ne w_p$. Summarising:

(1) the last two entries of $F(\tilde{y})$ are z_{k-1}, w_p;
(2) since these are distinct they are also the last two entries of $\rho(F(\tilde{y}))$.

Because the track-sum of $\omega_{\rho(F(\tilde{y}))}$ with γ is not a geodesic, w_p is not between z_{k-1}, w_{p+1}. Otherwise, appending w_{p+1} to $\rho(F(\hat{y}))$ gives an irreducible extreme, whose curve is a geodesic by Lemma 4.1. However the curve is the track-sum of $\omega_{\rho(F(\hat{y}))}$ with γ. So w_p does not lie in the image of a minimal geodesic from z_{k-1} to w_{p+1} and consequently

$$\tilde{\delta} \equiv d(z_{k-1}, w_p) + d(w_p, w_{p+1}) - d(z_{k-1}, w_{p+1}) > 0.$$

Now z_{k-1}, w_p, w_{p+1} are the entries in positions $k-1, k, k+1$ respectively of $y' \equiv G_{k-1} \circ G_{k-2} \circ \cdots \circ G_1(y)$. By Lemma 2.3,

$$\lambda(F(y)) = \lambda(G_{q-1} \circ G_{q-2} \circ \cdots \circ G_k(y')) \leq \lambda(G_k(y')) = \lambda(y') - \tilde{\delta} < \lambda(y)$$

since $\tilde{\delta} > 0$ and by Lemma 2.3 again. The third assertion is proved.

LEMMA 4.3. *$\omega_{s^{(\infty)}}$ is a geodesic.*

PROOF. By Lemma 2.4, and because λ and F are continuous

$$\lambda(s^{(\infty)}) \geq \lambda(F(s^{(\infty)})) = \lim_{j\to\infty} \lambda(F^{a_j+1}(y)) \geq \lim_{j\to\infty} \lambda(F^{a_{j+1}}(y)) = \lambda^{(\infty)}$$

again by Lemma 2.4. But $\lambda^{(\infty)} = \lambda(s^{(\infty)})$. Therefore $\lambda(F(s^{(\infty)})) = \lambda(s^{(\infty)})$ and the lemma follows from Lemma 4.2.

5. Geodesics between distant points

As in Section 2 let $\omega : [0, 1] \to N$ be a given piecewise-C^1 curve parameterized proportionally to arc-length. Construct $y \in Y$ from ω as in Section 2. Define $F : Y \to Y$ as in Section 2 and let $s^{(\infty)}$ be the limit of any convergent subsequence of $S = \{F^a(y) : a \geq 1\}$. Note that Y is compact so that at least one convergent subsequence exists. Let $\gamma : [0, 1] \to N$ denote the curve $\omega_{s^{(\infty)}}$ of $s^{(\infty)}$ defined as in Definition 3.2. Then γ is the limit of the subsequence $\{\omega_{s^{(a_j)}} : j \geq 1\}$ of $\Omega = \{\omega_{s^{(a)}} : a \geq 1\}$.

THEOREM 5.1. (1) *γ is a geodesic*;
(2) *if ω is already a geodesic then $\gamma = \omega$*;
(3) *$L(\gamma) \leq L(\omega)$ and $L(\gamma) < L(\omega)$ unless ω is a geodesic*;
(4) *γ is homotopic to ω through curves joining $\omega(0), \omega(1)$*;

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

(5) *Ω is uniformly convergent to γ, unless there exists a geodesic $\tilde{\gamma} \neq \gamma : [0, 1] \to N$ from x_0 to x_1 but homotopic to γ through curves joining x_0, x_1, and satisfying $L(\gamma) = L(\tilde{\gamma})$.*

PROOF. Lemma 4.3 says that γ is a geodesic. For every $z = F^a(y)$ we have $z_0 = y_0 = \omega(0)$, $z_q = y_q = \omega(1)$ and so this holds for the limit $s^{(\infty)}$ as well. Then from Definition 3.2 $\gamma(0) = \omega(0)$, $\gamma(1) = \omega(1)$.

If $L(\omega) = L(\gamma)$ then $L(\omega) = \lambda(y)$ because $L(\gamma) \leq \lambda(y)$. Then for $0 < i < q$ each $\omega|[t_{i-1}, t_{i+1}]$ has the same length as the minimal geodesic joining y_{i-1}, y_{i+1} and consequently is a minimal geodesic. Here we are using the hypothesis that ω is parameterized proportionally to arc-length. So $\omega = \omega_y$. By (4) of Section 3 $\omega_{F(y)} = \omega_y$. Arguing inductively, $\omega_{s^{(a)}} = \omega_y$ for all $a \geq 1$. Restricting attention to the convergent subsequence $\{s^{(a_j)} : j \geq 1\}$, $\gamma = \omega_y$. Therefore $\omega = \gamma$.

For $0 < i \leq q$ and $0 \leq u \leq 1$ replace $\omega|[t_{i-1}, t_{i-1} + u(t_i - t_{i-1})]$ by the minimal geodesic defined over the same subinterval and joining the same two points. Doing this for every i gives a homotopy from ω to ω_y through piecewise-C^1 curves from $\omega(0)$ to $\omega(1)$.

Any $z \in Y$ is also a point in the Cartesian power N^{q+1} which is a Riemannian manifold, complete with respect to the uniform metric d^{q+1}. By Lemma 2.5 there is a minimal geodesic from z to $F(z)$ whose image is entirely contained in

$$\tilde{Y} = \{w_0 \times \cdots \times w_q : w_0 = \omega(0), d(w_{i-1}, w_i) \leq 5\delta$$

for $0 < i \leq q$, $w_q = \omega(1)$.

So for $a = 0, 1, 2, \ldots$ let $\gamma_a : [0, 1] \to \tilde{Y}$ be the minimal geodesic from $s^{(a-1)}$ to $s^{(a)} = F(s^{(a-1)})$. Here $\gamma_0 : [0, 1] \to \tilde{Y}$. Choose j so large that $d^{q+1}(s^{(a_j)}, s^{(\infty)}) \leq 2\delta$ and let $\gamma_\infty : [0, 1] \to \tilde{Y}$ be the minimal geodesic from $s^{(a_j)}$ to $s^{(\infty)}$. A track-sum of these minimal geodesics gives a continuous path $h : [0, 1] \to \tilde{Y}$ from y to $s^{(\infty)}$.

The curve construction of Definition 3.2 applies also to elements of $\tilde{Y}$ because δ was chosen conservatively in Section 2. (If the homotopy is not required then δ can be taken twice as large in Section 2.) Applying the curve construction to each point on the continuous path h yields a homotopy from ω_y to γ through piecewise-geodesics joining $\omega(0), \omega(1)$.

To prove the last assertion suppose that there is no geodesic $\tilde{\gamma}$ with the properties listed. Then the limit $\tilde{s}^{(\infty)}$ of any other convergent subsequence of S gives rise to the same geodesic γ. Let

$$\Theta = \{\omega_y : [0, 1] \to N : y \in Y\}$$

with the quotient topology from Y, namely the topology of the uniform metric. Then Θ is compact and γ is the only accumulation point of the subset Ω. This completes the proof.

https://doi.org/10.1017/S1446788700039380 Published online by Cambridge University Press

COROLLARY 5.1. *Let N have everywhere non-positive sectional curvature. Then $\{\omega_a : a \geq 1\}$ is convergent.*

PROOF. By [4, Theorem 8.1] the exponential $\exp_{x_0} : TN_{x_0} \to N$ is a covering map and γ, $\tilde{\gamma}$ would lift to the curves $t \to t\dot{\gamma}(0)$, $t\dot{\tilde{\gamma}}(0)$ in TN_{x_0}. A homotopy from γ to $\tilde{\gamma}$ would lift to a homotopy in TN_{x_0}. But $\exp_{x_0}^{-1}(x_1)$ is discrete, and the contradiction proves the corollary.

References

[1] R. Bott, 'The stable homotopy of the classical groups', *Ann. of Math.* **70** (1959), 313–337.

[2] P. B. Chapman and J. L. Noakes, 'Singular perturbations and interpolation - a problem in robotics', *Nonlinear Anal.* **16** (1991), 849–859.

[3] H. B. Keller, *Numerical methods for two-point boundary-value problems* (Blaisdell, Waltham, 1968).

[4] S. Kobayashi and K. Nomizu, *Foundations of differential geometry, Volume II* (Interscience, New York, 1969).

[5] C. A. Micchelli, 'On a measure of dissimilarity for normal probability densities', preprint, IBM Yorktown Heights, 1996.

[6] J. W. Milnor, *Morse theory*, Ann. of Math. Stud. 51 (Princeton Univ. Press, Princeton, 1963).

[7] L. Noakes, 'Nonlinear corner-cutting', *Adv. Comput. Math.*, to appear.

[8] ———, 'Asymptotically smooth splines', in: *Advances in Computational Math., Series in Approx. Decompositions* 4 (World Scientific, Singapore, 1994) pp. 131–137.

[9] ———, 'Riemannian quadratics', in: *Curves and surfaces with applications in CAGD* (eds. L. L. Schumaker, A. Le Méhauté and C. Rabut) (Vanderbilt University Press, 1997) pp. 319–328.

[10] L. Noakes, G. Heinzinger and B. Paden, 'Cubic splines on curved spaces', *IMA J. Math. Control Information* **6** (1989), 464–473.

[11] R. S. Palais and C.-L. Terng, *Critical point theory and submanifold geometry*, Lecture Notes in Math. 1353 (Springer, Berlin, 1988).

[12] C. R. Rao, 'Information and the accuracy attainable in the estimation of statistical parameters', *Bulletin Calcutta Math. Soc.* **37** (1945), 81–91.

[13] L. T. Skovgard, 'A Riemannian geometry of the multivariate normal model', *Scand. J. Statist.* **11** (1984), 211–223.

[14] K. L. Teo, C. J. Goh and K. H. Wong, *A unified computational approach to solving optimal control problems* (Longman Scientific & Technical, 1991).

[15] J. H. C. Whitehead, 'Convex regions in the geometry of paths', *Quart. J. Math. Oxford* **3** (1932), 33–42.

[16] E. Zeidler, *Nonlinear functional analysis and its applications II/B* (Springer, Berlin, 1990).

[17] Z.-Q. Zuo, 'Two new techniques for optimal control', *IEEE Trans. Autom. Control* **36** (1991), 1307–1310.

Department of Mathematics
The University of Western Australia
Nedlands WA 6907
Australia
email: lyle@maths.uwa.edu.au